1. The journey of I.C.Technology from micro(1959) to nano era(2009)
    1. The Journey of I.C.Technology from micro (1959) to nano (2009) era_ABSTRACT.
    2. The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 1.
    3. The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 2.
    4. The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 3.
    5. The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 4.
    6. The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 5.
    7. The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 6.
    8. The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 7.
    9. The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 8.
    10. Part 9_Journey of IC
    11. 50 years Journey of IC Technology_Appendix I
    12. 50 years Journey of IC Technology_Appendix II
    13. 50 years Journey of IC Technology_Appendix III_Alternatives of NGL
    14. 50 years Journey of IC Technology_Appendix IV_Band Gap Narrowing and its effect on Emitter Injection Efficiency
    15. 50 years Journey of IC Technology_Appendix V_Polysilicon Emitter BJT
    16. 50 years Journey of IC Technology_Appendix VI +VII_Atomic Layer Deposition+FermiLevelPinnning
    17. 50 years Journey of IC Technology_Appendix VIII_Mobility Enhancement in strained Silicon
    18. 50 years Journey of IC Technology_Appendix IX_History of SiGe HBT

2. The Unfolding of Third Information Revolution
    1. Solid State Physics & Devices _The harbinger of third wave of civilization_PREFACE_Unfolding of Third Information Revolution.
    2. The onward march of Third Information Revolution in the first decade of 21st Century.
    3. Moore Law continues in the second decade of 21st Century.
    4. FREQUENCY SPECTRUM ALLOCATION IN 21st Century.
    5. The upgraded Internet Protocol IPv6 and challenges of expolsively growing INTERNET(Untitled)
3. SSPD_Chapter 1_Modern Physics-the foundation of Solid State Physics and Devices.
    1. SSPD_Chapter1_Modern Physics-Foundation of Solid State Physics_Part1
    2. SSPD Chapter 1_Part 1_Appendices of Disoverers and Inventors
    3. SSPD Chapter 1_Part 2_Photoelectric Effect
    4. (SSPD_Chapter 1_Part 2 continued
    5. Theory of Dualism as proposed by the French Prince de Broglie
    6. SSPD Chapter 1_Part 3_Theory of Dualism_Matter wave continued
    7. SSPD Chapter 1_Part 4_Heisenberg Uncertainty Principle
    8. Story of Atom_Plum-Pudding Model of Atom as proposed by J.J.Thomson.
    9. SSPD Chapter 1_Part 5_Story of Atom-Solar System Model of Atom as proposed by Rutherford.
    10. SSPD_Chapter1_Part6_Neil Bohr's Quantum Theory of Atoms.
    11. SSPD_Chapter1_Part 7_Hydrogen Atom Analysis
    12. SSPD_Chapter1_Part 7continued_Theoretical Formulation of Rydberg Constant
    13. Atomic Energy Exchanges and Physics of Gas Discharge Tubes.

14. SSPD_Chapter1_Part 8_Electron in a Potential well_Formulation of Schrodinger Equation.
15. SSPD_Chapter1_Part 8continued_Electron in an Infinite Potential Well_Solution of Schrodinger Equation.
16. SSPD_Chapter1_Part 8_continued_Electron in an Finite 1-D Potential Well_Solution of Schrodinger Equation.
17. SSPD_Chapter1_Part 8_continued_Electron in an 1-D Potential Well of finite height and finite width_Quantum Mechanical Tunneling
18. SSPD_Chapter1_Part 8_conclusion_Electron travelling across a step potential
19. SSPD_Chapter 1_Part8_AppendixXXXIV_Travelling Waves
20. SSPD_Chapter 1_Part8_AppendixXXXV_Schrodinger Eqaution and its solution
21. SSPD_Chapter 1_Part 9_PAULI-EXCLUSION PRINCIPLE
22. SSPD_Chapter 1_Part 9_Conclusion_Removal of degeneracy
23. SSPD_CHAPTER 1_Part10_ ELECTRON IN A SINGLE CRYSTAL SOLID AND BAND THEORY OF SOLID.
24. SSPD_CHAPTER 1_Part10_ Continued_The theoretical basis of Band Theory of Solids
25. SSPD_CHAPTER 1_Part10_ Concluded_Kronig-Penney Model
26. SSPD_Chapter 1_Part 11_Solid State of Matter
27. SSPD_Chapter 1_Part 11_Solid State of Matter_diifferent bondings
28. SSPD_Chapter 1_Part 11_Solid State of Matter_Summary of Bonds
29. SSPD_Chapter 1_Part 11_Solid State of Matter_Crystalline Nature of Solids
30. SSPD_Chapter 1_Part 11_Solid State of Matter_continued_Direct and indirect band-gap materials
31. SSPD_Chapter 1_Part 11_Solid State of Matter_continued_Crystal Structure of Si,Ge and GaAs.
32. SSPD_Chapter 1_Part 11_Solid State of Matter_Conclusin_Defects in Crystal Structures
33. SSPD_Chapter 1_Part 12_Quantum Mechanical Interpretation of Resistance

   34. SSPD_Chapter 1_Part 13_ Constituent particles of Solid State of Matter_Their Birth and Evolution.
   35. SSPD_Chapter1_Part 13_Conclusions of Big Bang.
   36. The Epic Quest has ended and the New Campaign has started
   37. SSPD_Chapter 1_Part 13_ Asymmetrical Universe
   38. SSPD_Chapter 1_Part 14_BOSONS , FERMIONS and Physics of LASER ACTION.
   39. SSPD_Chapter 1_Part 15_Quantum Fluids_Superconductivity & Superfluidity.
   40. SSPD_Chapter 1_Part 16_FERMIONS_Part 17_Density of States in Conduction Band of Metals
   41. SSPD_Chapter 1_Part 18_Kinetic Theory of Gases from QM perspective
   42. SSPD_Chapter 1_Appendices_Space Weather Research

4. SSPD_Chapter 2_Semiconductor Physics

   1. Crystal Growth-Bulk and Epitaxial Film-Part 1
   2. Crystal Growth-Bulk and Epitaxial Film-Part 2
   3. Crystal Growth-Bulk and Epitaxial Film-Part 3
   4. Crystal Growth-Bulk and Epitaxial Film-Part 4
   5. Syllabus_EC1410A_Solid State Physics and Devices in nano-technology era.
   6. History of Electron Devices- Vacuum Tube Era, Solid State Era and NanoTechnologyEra(Untitled)
   7. 2.1.2. Solid State Era
   8. SSPD_Chapter 2.1.3.Nano-Technology Era.
   9. Supplement_2.1.3.2.Transport in NanoTechnology Era.
   10. SSPD_Chapter2.2.Semiconductor Physics with emphasis on Silicon
   11. SSPD_Chapter2.2.1.and 2.2.1.Solid State Equivalent of Vacuum Devices and Crystalline structure of Silicon
   12. Chapter2.2.3;2.2.4;2.2.5;Insulator,Semi-conductor,Metal;Intrinsic Semi-condctor;Extrinsic Semi-conductor.
   13. SSPD_Chapter_2.2.6. Theoretical Formulation of thermal equilibrium values of majority and minority carriers in Semi-conductors.
   14. SSPD_Chapter_2.2.6. Drift Velocity in Semiconductor/

    metal and its conductivity.
15. SSPD_Chapter_2.2.9-2.2.10-2.2.11:Diffusion,Life-Time and Continuity Equation.
16. Tutorial Sheet of Chapter 2.Semiconductor Physics.

5. SSPD_Chapter 3_Diode Physics

    1. SSPD_Chapter 3_Solid State Diode Physics.
    2. SSPD_Chapter 3_Section 3.2.3.Theoretical Formulation of the Knee Voltage(Vγ) of a forward biased-diode.
    3. SSPD_Chapter 3_SEc3.2.4. Reverse biasedDiode,3.2.5.Real Diodes,3.2.6.Avalanche Diode,3.2.7.Hyperabrupt Jn.Diode.
    4. SSPD_Chapter 3.2.7.1.. Application of Hyperabrupt Junction Diode as Varactor Diode in Voltage-controlled Oscillator(VCO).
    5. AnalogElectronics_Lecture3_Incremental Model of Diode.
    6. AnalogElectronics_Lecture3supp._High Frequency Model of PN Junction Diode.
    7. SSPD_Chapter 3_Section3.3. Quantum Mechanical Diodes.
    8. Chapter 3_Section3.3.3Tunnel Diodes,3.3.4.Backward Diodes
    9. SSPD_Chapter 3_Section 3.4.Schottky Diode and Section 3.5. Ohmic Contact.
    10. Tutorial Sheet of Chapter 3.Diode Physics.
    11. SSPD_Experiment 2_Figures

6. SSPD_Chapter 4_Transistor Physics

    1. SSPD_Chapter 4_Bipolar Junction Transistor Physics
    2. Analog Electronics Lecture 2_PartB_BJT configurations & modes
    3. Analog Electronics Lecture 2_PartC_I-V output characteristics of BJT
    4. AnalogElectronics_Lecture4_PartA_dc,low&high frequency model of CB BJT.
    5. AnalogElectronics_Lecture4_PartB_low and high frequency model of CE BJT.
    6. SSPD_Chapter 4_ Section 4.1.Physical Significance of

    hybrid parameters (h-parameters) of BJT.
    7. Chapter 4. Section 4.2. Structure of IC NPN BJT.
    8. Chapter 4_Section 4.3. Design Rules for fabricating BJT.
    9. Chapter 4. Section 4.3.supplement-Solved Problems on Design of BJT.
    10. AE Lecture 8_Part A_MOSFET & JFET continued
    11. AE_Lecture8_MOSFET & JFET_PartA
    12. AE_Lecture8_PartB_Incremental Model of FET
    13. Chapter 4_Section 4.4._Design Rules for Fabricating nJFET.

7. SSPD_Chapter 5_Quantum Optical Devices_Photonics

    1. SSPD_Chapter 5_Section 5.1. Photo-Diodes and its Responsivity Curves.
    2. Chapter 5_Section 5.2. Direct and in-direct Band Gap
    3. SSPD_Chapter 5_Section 5.3. Spectal Responsivity of Photo-Diodes.
    4. SSPD_Chapter 5_Section 5.3_Supplement_Photo-Diode_I-V curve.
    5. SSPD_Chapter 5_Section5.5._Photo Sources.

8. SSPD_Chapter 6_Design and Fabrication of IC Chips_Basics

    1. Syllabus of EC1X61_IC Design and Fabrication Technology-Microelectronics [3-1-0]
    2. Chapter 6. IC Fabrication Technology from 60s to date_Part 1
    3. Chapter 6. IC Fabrication Technology from 60s to date_Part 2
    4. SSPD_Chapter 6_Part 3_The challenges of IC Technology
    5. SSPD_Chapter 6_Part 3_concluded_The list of Foundaries
    6. SSPD_Chapter 6_Part 4_The instability of MOS in 60s and its soluion
    7. SSPD_Chapter 6_Part 4_concluded_MOS Physics
    8. SSPD_Chapter 6_Part 5_Crossing of the Rubicon_Emergence of MOS Technolgy as the dominant IC Technology.
    9. SSPD_Chapter 6_Part 6_MOS ICs Fabrication Technology_to be continued
    10. SSPD_Chapter 6_Part 6_MOS ICs Fabrication

    Technology_to be continued2
    11. SSPD_Chapter 6_Part 6_MOS IC's Fabrication Technology_To be continued3
    12. Conclusion of CMOS Flow Process
    13. Class Test 1_25th_March_2011
    14. SSPD_Chapter 6_Part 7_Device Simulation1
    15. SSPD_Chapter 6_Part 7_Device Simulation2
    16. SSPD_Chapter 6_Part 7_Device Simulation2_concluded
    17. SSPD_Chapter 6_Part 7_Device Simulation2_Tony Plots
    18. SSPD_Chapter 6_Part 7_Device Simlation 3
    19. SSPD_Chapter 6_Part 7_Device Simlation 3_conclsions
    20. SSPD_Chapter 6_Part 7_ Diode Tutorial
    21. SSSPD_Chapter 6_Part 7_Introduction to ATHENA_Process Simulator
    22. SSPD_Chapter 6_Part 7_Introduction to ATHENA 2_Process Simulator
    23. SSPD_Chapter 6_Part 7_Introduction to ATHENA3_Process Simulator
    24. SSPD_Chapter 6_Part 7_Introduction to ATHENA 4_Process Simulator
    25. SSPD_Chapter 6_Part 7_Introduction to ATHENA 5_Process Simulator
    26. SSPD_Chapter 6_Part 7_Introduction to ATHENA 6_Process Simulator
    27. SSPD_Chapter 6_Part 7_Introduction to ATHENA 7_Process Simulator
    28. SSPD_Chapter 6_Part 8_Caliberating ATHENA for typical MOSFET FLOW
    29. SSPD_Chapter 6_Part 9_Caliberating ATHENA for typical Bipolar Flow Process
    30. SSPD_Chapter 6_Part 9_Appendix Device Physics
    31. SSPD_Chapter 6_Part 9_Figure IX of Appendix on Device Physics
    32. SSPD_Chapter 6_Part 10_Simulation of MOSFET
    33. SSPD_Chapter 6_Part 10_Simulation of MOSFET_conclusion

9. SSPD_Chapter 7_VLSI Design of CMOS Digital Systems.

    1. SSPD_Chapter 7_Syllabus of CMOS DIGITAL VLSI

DESIGN

2. SSPD_Chapter 7_Part 1_The Art of Microchip Fabrication transforms into HiTech Computer-Aided VLSI DESIGN and Microchip Fabrication.(Untitled)
3. SSPD_Chapter 7_Part 2_BiCMOS Technology
4. SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOS and MOS circuits to be continued1
5. SSPD_Chapter 7_Part3_Electrical properties of MOS continued2
6. SSPD_Chapter 7_Part 3_Electrical Properties of MOScontinued3
7. SSPD_Chapter 7_Part 3_Basic Electrical Properties_continued4.
8. SSPD_Chapter 7_Part3_Basic Electrical Properties of MOScontinued5.
9. SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOScontinued6
10. SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_continued7
11. SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_continued8
12. SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_concluded.
13. SSPD_Chapter 7_part 4_CMOS lay-out design rules.
14. SSPD_Chapter 7_part 4_CMOS lay-out design rules_continued2.
15. SSPD_Chapter 7_part 4_CMOS lay-out design rules_concluded.
16. SSPD_Chapter 7_Part 5_Stick Diagram of Logic Gates.
17. SSPD_Chapter 7_Part 5_Stick Diagrams of Logic Gates_continued 2
18. SSPD_Chapter 7_Part 5_Stick Diagram of CMOS Logic Gates_continued 3
19. SSPD_Chapter 7_part 5_Stick Diagrams of CMOS Gates_concluded.
20. SSPD_Chapter 7_Part 6_Basic Circuit Concepts.
21. SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued2
22. SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 3

    23. SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued4
    24. SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 5
    25. SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 6
    26. SSPD_Chapter 7_Part 6_Basic Circuit Concepts_concluded
    27. SSPD_Chapter 7_Part 7_Scaling of MOS circuits.
    28. SSPD_Chapter 7_Part 7_Scaling of MOS circuits_continued2
    29. SSPD_Chapter 7_Part 7_Scaling of MOS circuits_concluded.
    30. SSPD_Chapter 8_The Integrated Circuit Technology of Future Generation

        1. The Integrated Circuit Technology of Future Generations-Part 1.
        2. Graphene IC Part 2_Graphene IC + Graphene Allotropes
        3. I.C.Chips of Future Generation Part 3.carriers-phonon interaction in Graphene.
        4. The Integrated Circuit Technology of Future Generations-Part 3_mobility measurement, defect estimate, balllistic nature, conductivity variation.
        5. The Integrated Circuit Technology of Future Generations-Part 4_QED properties continued
        6. The Integrated Circuit Technology of Future Generations-Part 5_QED properties continued

10. SSPD Lab Experiments

    1. Experiment no -1: Components Testing
    2. SSPD-Experiment Number 2
    3. SSPD_Experiment2_Frequency Domain Response of HPF
    4. SSPD_Experiment 2_Time Domain Response of HPF.
    5. SSPD_Experiment 2_TDR of HPF_Part2
    6. SSPD_Experiment 2_Figures of LPF_Part1
    7. SSPD_Experiment 2_Time Domain Response of LPF
    8. SSPD_Experiment 2_Time Domain Response of

LPF_Part2

The Journey of I.C.Technology from micro (1959) to nano (2009) era_ABSTRACT.

Summary: This is the abstract of the main article named as above.

The Journey of I.C.Technology from micro (1959) to nano (2009) era_ABSTRACT.

Keyword: Monolithic Planar Technology, Wafer, epitaxial layer, fabrication, photolithography;

Summary: This is the abstract of the main article named as above.

ABSTRACT:

The discovery of ELECTRON by J.J. Thomson started a series of inventions which was to culminate into Solid State Bipolar Junction Transistor (invented by Shockley, Bardeen and Brattain) in the year 1948. This marked the advent of the era of *Solid State Technology*. Solid State Technology laid the foundation of *Silicon Monolithic Planar Integrated Circuit Technology* in 1959. In 1960 first high temperature epitaxial based BJT IC came into market. This kick started a process of *Information Revolution* leading to the third wave of civilization or post industrial civilization. By 1970 four- bit Microprocessor Chip uP4004 was introduced. The advancement of IC Technology led to exponential increase in packing density as well as

in electrical performance. From 1971 to 2001, *Small Scale Integrated* Chips evolved to *Medium Scale Integrated* Chips to *Large Scale Integrated* Chips to *Very Large Scale Integrated* Chips to the present Pentium IV which is *Ultra Large Scale Integrated* Chips packing 42 million CMOS Chips. This increase in packing density was enabled by the advancement the lithography technique. In 1966 emulsion mask patterns were contact printed on 200mm wafer with the smallest feature size of 25μm. In 2007 using Deep Ultra Violet(DUV)193nm immersion Lithography, 30nm ±6nm feature size are imprinted on 300mm wafer. We moved from Wet Processing to *Dry Processing*(reactive ion etching) or *Ion beam Processing* , Diffusion was replaced by *Ion Implantation*, Mask Making Techniques were improved and automated, preparation of master art work was automated using computer, step and repeat cameras were replaced by contact cameras, photolithography used ultra violet light for exposing the Kodak Photo Resist, UV light lithography was replaced by *Deep UV lithography* and now it is in process of being replaced by *Extreme UV laser beam lithography* as the feature size became smaller, Liquid Phase Epitaxy and Chemical Vapour Phase Epitaxy developed into *Metal Organic Vapour Phase Epitaxy*, new materials were introduced for interconnections (such as *silicides* and *copper*) and *Poly- Silicon* for gate contacts, the dielectric material used in gate silicon dioxide is being replaced by *hafnium oxide* which has a higher dielectric strength and higher k , planar technology is being replaced by *trigate* technology where gate contact is on the top as well as on the side walls of the oxide layer thereby increasing the contact area, parasitic were reduced by shallower junctions, smaller feature size and oxide side walls and by using insulator as the substrate for instance silicon on sapphire. One more innovation was *Lightly Doped Drain* for improving the reliability of the MOS devices. FET depends on leading edge lithographic dimensions(45nm) but BJT depends on vertical base widths which has reached 10nm. BJT has structural flexibilities in minimizing the parasitic. It has higher trans-conductance, high self gain, low 1/f flicker noise and better V_{BE} matching. All these factors make BJT the device of choice for demanding applications.While microlevel BJT scaled down to nano level structure, new problems arose due to shallow EB junction, band gap narrowing in emitter due to doping concentration exceeding above 10^{18} / cc and high sheet resistance of the base layer. These problems were tackled by using *heavily implanted PolySi layer used as emitter contact*. *SiGe Hetro Bipolar Transistors(HBT)* hold the key to realizing high speed wire-line and wireless communication circuits and systems. Marriage of Si Technology and Bandgap Engineering Methods of III-V Compund Semiconductor which is broadly called SiGe based bandgap engineering is Si Heterostructure Bipolar Technology (SiGe HBT). This was made possible by Low Temperature Epitaxy(LTE). Scaling strategy was used here also for exponential growth in the packing density and electrical performance of devices. To reduce the time delays, base width and collector widths have to be reduced and I_C must be increased which requires Kirk Effect should be pushed to higher Current Densities. This is achieved by increasing Collector Dopent Concentration. Hence *Selectively Implanted Collector* (SIC) is introduced. Decreasing the distance between SIC and extrinsic base implant will increase the overlap i.e. C_{CB}OL. This has to be minimized. This is achieved by *Raised Extrinsic Base* structure. An optimization in Ge profile in SiGe base layer has to be done inorder to get the full advantage of increased base doping in HBT. Today SiGe HBT are surpassing GaAs HBT. The integration of SiGe HBT with CMOS is BiCMOS. BiCMOS reduces the cost of mobile consumer products, advance high BW wireless communication and collision-avoidance automobile radar. BiCMOS lead to VLSI, ASIC & Si based RF SOC(System on Chip) solution. Hence SiGe/SiGeC and BiCMOS is becoming the technology of future. The strategy of *Sidewall oxidation* and *Silicon-on-Insulator*(SOI) is adopted for reducing the parasitic capacitances and improving the frequency response. P+ isolation diffusion is replaced by trench isolation. *Shallow Trench Isolation*(STI) is achieved by Reactive Ion Etching(RIE). This considerably helps reduce the parasitic and helps improve the speed. Since 90-nm Technology generation was introduced, off-state leakage current and power density in CMOS have made scaling a difficult and challenging job. New scaling vectors were adopted to meet this challenge. The new scaling vector was *Mobility enhancement through strained Si* and novel structure like FINFET . At 45-nm Technology and beyond, new flow process adopted was *High k + Metal Gate using Gate-Last strategy*. In the future we hope to go from 45nm technology to 32nm technology to 22nm technology to 16nm technology. Today in Bipolar Technology we have achieved 200GHz transit frequency and in CMOS Technology quad core microprocessor has come in the market. Intel and AMD have launched a native quad processor using 45nm technology. Both achieve the 4 cores on the same silicon slab. This has Memory Bandwidth. These conserve energy by cutting down power when the device is idle. These four cores are built up of 40billion transistors.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 1.

Summary: This is Part 1 of the main lecture on 50 years IC journey.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 1.

Keywords: Solid State Technology, PhotoVoltaic Cell, Cat’s Whisker, Single Crystal Silicon;

Summary: This is Part 1 of the main lecture on 50 years IC journey.

INTRODUCTION

Einstein had once remarked “ We stand on the shoulders of Giants”.

The experiences of the Savage Age laid the foundation of the Barbarian Age. The experiences of Barbarians laid the foundation of Agricultural Civilization. The Guild Houses of the Agricultural Society and the Maritime Traders with their mercantile capital paved the way for the Industrial Society.So to understand and appreciate the Integrated Circuit (IC) Technology evolution we must go to 18th century during the renaissance sweeping the Western World. Table I.1 (Appendix I) gives the road map of development of Solid State Technology , Table I.2 (Appendix I) gives the road map of development of Electronics and Integrated Circuit Technology and Table I.3 gives the development in Super Computers.

In early eighteenth century, metal as a conductor and its positive temperature coefficient of resistance was recognized. This was followed by the discovery of a class of solids which exhibited negative temperature coefficient of resistance. The passage of current caused thermal runaway increase in current. This gave birth to Semi-Conductors. Contact potential between dissimilar materials was observed. In a series of experiments photo-voltaic effects in Ag-AgCl interface and light dependent resistance in Selenium was observed. This was followed by the construction of PV-cell and point contact diode. Point contact diode was fabricated by pressing a wire-point on Galena(PbS). This was also known as cat's whisker. The cat's whisker replaced the coherer in Radio Receivers. J.C. Bose patented cat's whisker diode and AT&T patented a Silicon Crystal diode. By early 20th century point contact diode was widely used as diode peak detector in Radios. This point contact diode was replaced by Vacuum tube Diode. CuO-Se was developed as PV Cell.

In 1930 the theoretical basis for Solid State Technology was laid with the publication of " Quantum Theory meets Semiconductors" by Alan F. Wilson. This book contained the Band Theory of

Solids, thermal generation of electron-hole pairs and the concept of intrinsic and extrinsic semiconductor. Solid State Processing led to the preparation of Electronic Grade Polycrystalline Silicon.

At the end of 1947, point contact Ge- Transistor was invented which behaved like pentodes by Shockley , Bardeen and Braittain. Its power gain was tested by building an oscillator. But point contact Ge-Transistor were unstable, unreliable and became non-functional above 75ºC. Hence a transition was made to single crystal diffused Silicon Transistor.

The first npn BJT made of pure single crystal Silicon at BELL Labs by rate growing technique (made by Morris Tanenbaum and announced at June 1954 IRE Solid State Device Research Conference. The same paper was published in peer reviewed Journal of Applied Physics in June 1955).

1. Samples bought from DuPont.
2. A high purity single crystal Si ingot grown from the sample.
3. 1.27cm length bar was sliced.
4. By rate growing technique narrow base region was implemented (13 to 25 μm).
5. On 26th January 1954, testing was completed.

In 1955 The first npn BJT made of pure single crystal Silicon at Texas Instrument.

(Announced at IRE conference, Dayton, Ohio).

1. High purity silicon bought from DuPont at $500 a pound.
2. 1.27cm length bar was sliced from the ingot.
3. P Type Base was 25μm thick- thin enough for minority carriers to diffuse through and get collected by the Collector..
4. N Type Emitter was heavily doped to give high injection efficiency.

Texas Instrument had Si BJT market to itself for rest of the decade and thus the stage was set for the emergence of TI as the International giant in field of device manufacturing.

At the time Telephone was invented in 1876 it could not become a Public Utility Service. Only after the invention of RC-coupled Amplifier using Vacuum Tube Triodes and Pentodes that long distance calls across the continents could be made. But even then scope of Electronics Amplifier and scope of Electronics Computer remained limited to Telecommunication and Defense Applications. This was on account of large physical sizes and large energy consumption. Solid State Transistors and later IC Technology only could overcome this limitation of Electronic Applications.

In 1959. Jack Kilby of Texas Instrument and Robert Noyce of FairChild almost simultaneously invented

Integrated Circuit Technology. In 1961 the first integrated RTL circuit was commercialzed and in 1964 first Analog IC MC 1530 Operational Amplifier was marketed.

In 1964 the third generation Computer using IC came into market. This was known as Table-Top Computer PDP-8 . This was compact. There was much less power requirement and much more cost affordable.

In 1971 the first microprocessor chip (μP4004) was marketed by INTEL . This laid the foundation of Desk-Top computers belonging to Forth Generation Computers.

The continuous scaling and innovation , led to the rapid micro-miniaturization and functionally complex IC chips as laid down according Moore's Law.

Gordon Moore predicted that number of transistors on integrated circuits(a rough measure of computer processing power) will double every 18 months at a minimum cost. It became a self fulfilling prophecy. Moore's Law has become a yardstick of our progress as we harness the cunning of NATURE's design strategies.

μP4004 was followed by 8085, 8086, 80386, 80486, Pentium I, Pentium II, Pentium III and Pentium IV. Today in 2009 such powerful quad processors have been commercialized that Nvidia advertises its new workstation, the Tesla, as a “personal supercomputer”. It clusters 4 Nvidia C1060 processing boards, each of which unites 240 graphic cores to process instructions at nearly teraflop speed. The evolution of microprocessors chip is given in Table I.5.

EMERGENCE OF KNOWLEDGE BASED SOCIETY.

At the beginning of Civilization we had agrarian economy and it had Labour Intensive Production Methods. With Industrial Revolution in 1750’s we had Capital Intensive Production. Today Computer and World Wide Web has created Knowledge Intensive Production. Today Computerization, Automation , Robotization and on-line Commerce has become imperative to face and survive the competition in the global market. After Industrial Revolution, Steel became the staple of society. In Post-Industrial society Silicon has become the staple of society. Following Table 1 shows the commanding position which Electronics Industry occupies in the present world economy. In fact Electronics Industry has become the main engine of growth for the global economy.

Table 1. Total chips production, electronic hardware generated and total services generated.

Total IC Chips & Discrete Components produced in 2004	Electronic Hardwares generated	Software services generated	Percentage of total world GDP
$213 billion	$1200 billion	$5000 billion	10%

As seen from Table 1, the Solid State Devices Industries and its spin off comprises 10% of World GDP(Total World GDP in 2009 is $50 trillion). With passage of time it is growing by leaps and bounds. Educated Guess of Gordon Moore is that 1018 Transistors(BJT & MOS) are produced annually. This is more than all the printed characters in all the newspapers. Today we sell the transistors cheaper than a character sold by Sunday New York Times. Such is the all pervasiveness of Transistors.

Today IC Technology has enabled the World Wide Web where 1 billion PCs are connected to the INTERNET sharing their data, sending and receiving emails and doing distributed computing with the help of hundred thousands of PCs connected to the NET. Google Search engines and their likes have made all data bases available to the NET users.

Zarlink Semiconductors has recently introduced M2A capsule endoscopy. These can travel down through the digestive tract and take photographs of two high-resolution colour images per second

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 2.

Summary: This Part 2 of 50 yrs IC journay.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 2.

Title: Part 2 of 50 yrs journey.

Keywords: Isolation Diffusion, base stripe geometry, buried layer, vertical scaling and lateral scaling.

Summary: This Part 2 of 50 yrs journay.

Section I. The Bipolar Junction Transistor Technology in 60’s-70’s.

In 1959 at the time IC Technology was invented, the vertical NPN transistor looked as given in Figure 1.

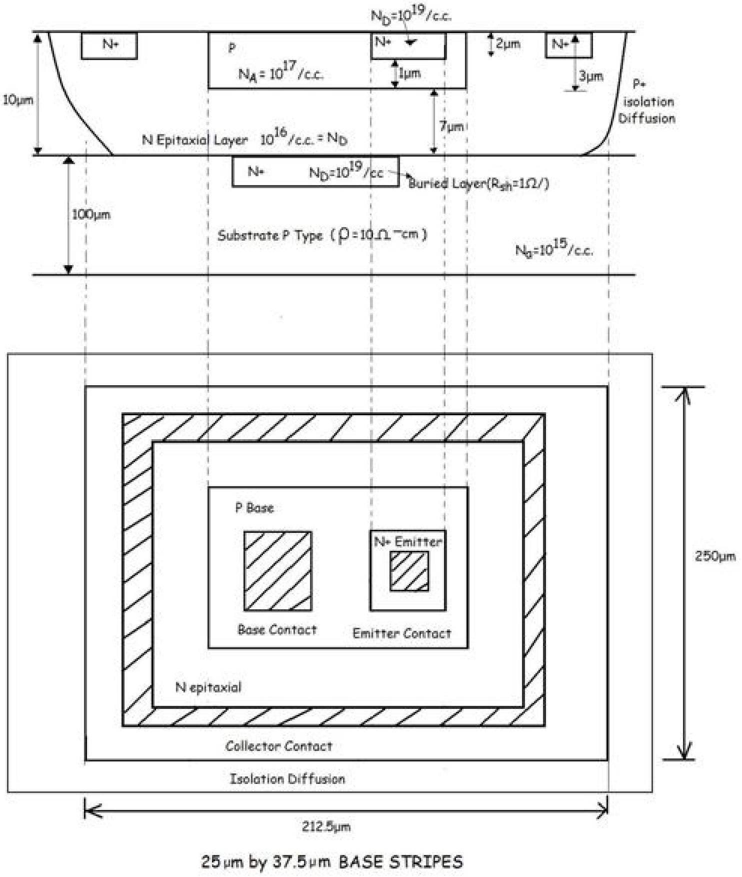


Figure 1. Cross-sectional View and plan view of Vertical NPN transistor.

(25um by 37.5um base stripe geometry)

Total area covered 212 μm **by 250** μm

Substrate is 100um thick.

Epitaxial layer is 10um thick.

Base Collector Junction is at 3um depth and Emitter-Base Jn. at 2um depth.

Hence Base Width = W_B = 1 um thick.

P Substrate- $\rho = 10\Omega$-cm, $N_A = 10^{15}$ Boron atoms/cc;

Buried Layer has been put to reduce Collector Series Resistance.

$N_D = 10^{19}$Arsenic Atoms/cc; In buried layer Arsenic is used as donor type dopent because it has much lower diffusion coefficient hence almost negligible out diffusion in subsequent diffusion and heating cycles.

N Epitaxial layer- $\rho = 0.1\Omega$-cm, $N_D = 10^{16}$Phosphorous Atoms/cc;

Base layer- $N_A = 10^{17}$ Boron Atoms/cc; Base layer has a sheet resistance of 200Ω/sq.

Emitter and Collector Contact layer- $N_D = 10^{19}$ Phosphorous Atoms/cc, R_{sh} = 1Ω/sq;

Section II. The Bipolar Junction Transistor Technology in 1980’s-2000’s.

Gordon Moore is a co-founder of Intel and he made an empirical observation which became a Law . It stated :

“ at our rate of technological development, the complexity of an integrated circuit, with respect to minimum component cost will double in about 24 months”.[“Cramming more components onto Integrated Circuits”, *Electronics Magazine,*19 April 1965]. The law has held the test of time to date as is evident from Table I.4.

Today Moore’s Law has become a self fulfilling prophecy and a goal Industry tries to achieve all the time. In attempting to maintain this rate of growth enormous R&D has gone into the development of tools and equipments required for IC fabrication and into the advancement of processing techniques required for IC fabrication.

IC Fabrication has two major parts:

1. System/circuit design and I.C. layout of the components;

2. Photolithographic Masks preparation

3. Actual fabrication of the IC chip.

The actual fabrication consists of the following steps:

1. Preparation of single crystal Silicon Wafer;
2. Buried layer diffusion in the substrate;
3. Epitaxial layer growth on the P Type substrate;
4. Isolation Diffusion (not required in CMOS

fabrication);
5. Base diffusion;
6. Emitter and Collector Contact diffusion
7. Metallization;
8. Interconnection;
9. Mounting on the header ,wire bonding by thermo-compression or ultra-sonic bonding and hermetical packaging;

During fabrication, suitable windows have to be etched in SiO_2 layer grown over the wafer. By the use of optical lithographic techniques, proper patterns are etched out using a suitable mask for selective diffusion or selective ion-implantation. This step is repeated at buried layer diffusion/ implantation, at isolation diffusion, at base diffusion, emitter & collector contact diffusion and at metallization step.

Obeying Moore’s law requires continuous scaling of the I.C. components and hence working at smaller and smaller feature size at the minimum cost. The main burden of this challenge falls on Mask Making and Lithography Techniques. Therefore the continuous and relentless march of IC from micron dimensions to nano dimensions has meant continuous improvement and breakthroughs in Lithography Technology. *In fact lithography technology is the major enabling technology in IC Chips fabrication.*

In Table 2 , Table 3 and Table 4 we show how the dimensions have been scaled down over the past decades in both Bipolar and MOS Technology. In 1966 emulsion mask patterns were contact printed on 200mm wafer with the smallest feature size of 25μm. In 2007 using Deep Ultra Violet(DUV)193nm immersion Lithography, 30nm ±6nm feature size are imprinted on 300mm wafer.

Table 2. Transistor Junction Depths and Doping Changes with years. [Muller & Kamini, Solid State Devices; “Modelling of SiGe heterojunction transistors:200GHz frequencies with symmetrical delay times”, Jochen Eberhardt & Erich Kasper, Solid State Electronics, 45(2001), 2097-2100.] .

	Amplifying(Junction Isolated)1964	Amplifying(Junction Isolated)1980	Switching(Junction Isolated)1980	Switching(Oxide Isolated)1980	Amplifying(oxi Isolated)2001H
Surface area	212.5μm × 250μm				20 μm × 1μm
Epi Film Thickness Resistivity of epi layer	10 μm0.1Ω-cm	10μm1Ω-	3μm0.3-0.8 Ω-cm	1.2um0.3-0.8 Ω-cm	17-28nmWc= 50nm
Buried Layer	1Ω/□2.5 μm	20Ω/□2.5μm	20Ω/□1.4μm	30 Ω/□0.3 μm	0.7 Ω/□0 μm

Sheet R_{Sheet} Updiffusion					
Emitter Diffusion Depth Sheet Resistance of Emitter	2.5 μm 1Ω/□	2.5μm 5Ω/□	0.8μm 12Ω/□	0.25μm 30Ω/□	W_E = 110nm 378 Ω/□ Emitter contact 5.2 Ω/□
Base Diffusion Depth R_{SHEET}	3.5 μm 200Ω/□	3.25μm 100 Ω/□	0.3μm 200 Ω/□	0.5μm 600 Ω/□	22nm 260 Ω/□
Base Width	1 μm	1 μm	0.5 μm	0.25 μm	W_B = 12nm
Substrate Resistivity Crystal Orientation	10 Ω-cm < 111 >	10 Ω-cm < 111 >	10 Ω-cm < 111 >	5 Ω-cm < 111 >	

Table 3. Dimension Scaling in MOSFET over the last decade.

MOS	1967	1997	1999	2001	2003	2006
L(μm)	10	0.25	0.18	0.13	0.1	0.07
DRAM **(Gbit/cm2)**	64M	0.18	0.38	0.42	0.91	1.85
Junction Depth **(xj)nm**	1000	100	70	60	52	40

Interconnection Pitch (nm)	2000	600	500	350	245	130

International Technology Roadmap for Semiconductors uses ‘nodes’ to refer to the smallest feature size on Logic Chips. This is the width of the gate of MOSFET. Whereas in Memory Chips ‘half pitch’ is used to define the smallest feature size. Half Pitch is half the distance between two adjacent aluminum interconnections. Nodes and Half pitch are illustrated in Figure 2.

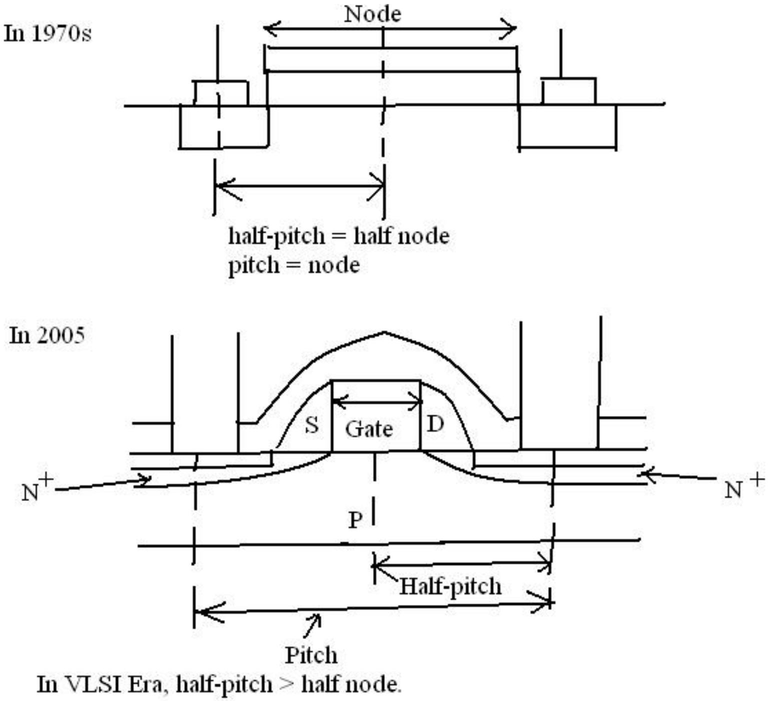


Figure 2. Illustration of the definition of node and pitch.

In 2005, gate width is 32nm, node is 65nm and half pitch is 105nm.

In 2009:

Memories are 1 full chip generation ahead of Logic chip.

Table 4. Scaling in IC Chips in terms of half pitch and nodes.

Year	Node(nm)	Half-pitch(nm)	Gate-length*(nm)
2009[a]	32	52	29
2007[a]	45	68	38
2005[b]	65	90	32
2004[b]	90	90	37
2003[b]	100	100	45
2001[c]	130	150	65
1999[c]	180	230	140
1997[d]	250	250	200

1995 [d]	350	350	350
1992 [d]	500	500	500

1. ITRS data 2008 update.
2. ITRS data 2006.
3. ITRS data 2001.
4. ITRS data 1997.

* The physical length has become smaller than printed length.

As feature size has reduced, shorter wave length has to be used hence light source in Lithography will have to be changed. If the printing feature is smaller than wavelength then light tends to get diffracted and edges get blurred. By image enhancement techniques, smaller feature size can be achieved with larger wavelength.

Now we will examine how Lithography Technology has evolved over the last 50 years.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 3.

Summary: This gives Part 3 of 50 yrs journey.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 3.

Title: Part 3 of 50 yrs journey.

Keyword: Scale of Integration, Contact printing, Projection printing, Proximity printing,UV, Deep UV,Immersion Technique.

Summary: This gives Part 3 of 50 yrs journey.

Section III. The Different Generations of Lithographic Technology serving the IC Fabrication Foundries over the last 50 years.

[“New Uses of IC Technology will enable better opportunities for optical lithography”, Resor & Griff, Solid State Technology,01-Nov07; “Technology Interdependence & the evolution of Semiconductor lithography,” Ron Adner & Rahul Kapur, Solid State Technology, Nov.2007; IC knowledge.com]

At the time I.C. technology was born mask preparation and optical lithography was manual[Appendix II].

As we move from Small Scale Integrated (SSI)

Circuits to Medium Scale Integrated Circuits(MSI) to LSI to VLSI and finally now to ULSI, the need for smaller and smaller feature size in circuit fabrication has driven the evolution of Optical Lithography. Table 5 shows the growth in the packing density of components in the chips fabricated in last 50 years.

Table 5. Classification of Chips according to degree of integration.

SSI(Small Scale Integrated Circuits)	Less than 30 components
MSI(Medium Scale Integrated Circuits)	30 to 100 components
LSI (Large Scale Integrated Circuits)	100,000 components
VLSI(Very Large Scale Integrated Circuits)	1 million components
ULSI(Ultra Large Scale Integrated Circuits)	40 million components

Table 6. The Development in Lithography Technique in last 50 years.

[“Shrinking Possibilities”, Bill Arnold, *IEEE*

Spectrum, April 2009, pp.23-25, 46-48.]

Year	Generation	W* (nm)	λ (nm) Min	Source	Year of dominance	Type of mask	Method of exposure	NumericalA
1962	Contact printing	7µm	436	UV	1972	Emulsion plates	Refraction through lens system	0.3

Late 60s	Computerization and automation of mask preparation using high speed step and repeat cameras. This gave reduced geometries and

	greater accuracy.							
1972	Proximity Printing	3µm	480	UV	1973		do	0.3
1973	Projection	2µm	480	UV	1977	Electron beam systems used in mask making	do	0.3
1976	Electron beam	0.5 µm	480	UV	never	.	Electrostaticlensin	0.3
1978	X-Ray	0.3 µm	480	UV	never	No suitable mask materials	?	0.3
1978	G-line	1.25 µm	436	G line of Hg vap.lamp	1982		Lens system.Refractive exposure	0.3
1985	I-line	0.80 µm	436/365	I line of Hg	1991		do	0.4
1986	DUV248 nm	0.45 µm	365	Same	1998		do	0.4
1996	DUV193nm	0.35 µm	365/248	KrF Excimer	2006		do	0.6
1998	DUV157nm	0.10 µm	248	do	Never		do	0.6
1999	DUV193nm	0.13 µm	248	do			do	0.6
2001	DUV193nm	0.13	248	do			do	0.7

2003	DUV 193nm	0.09 µm	248/193	ArF Excimer			do	0.8
2005	DUV 193nm immersion	0.04 µm	193	do	To date		do	0.8
future	EUV 100nm	Less than 0.030 µm					Reflective mirror exposure system**.	

W* = minimum feature size = (according to Rayleigh Equation) = $k.\lambda/(NA)$

where NA(numerical aperture) = $n.\mathrm{Sin}(\alpha) = d/(2f)$;

n = index of refraction of the medium surrounding the lens and α = acceptance angle of the lens system;

d = aperture diameter and f = focal length.

k = resolution factor. Resolution factor is also falling. It depends on the photo resist and resolution enhancement techniques such as :

phase shift mask,

off axis illumination

and optical proximity corrections (this corrects for the blurring effects of diffraction of light from small aperture slits). In the mask a rectangle pattern is replaced by dog bone pattern so that after exposure we get a rectangle with sharp corners on the wafer.

Increased NA gives reduced depth of focus. For high resolution, surface of Photoresist must be flat and within the depth of focus.

By looking at the Rayleigh Equation we immediately see the advantage of using immersion technique. The most suitable medium is non-ionized ultra pure water which has n = 1.47 and absorption less than 6% over 6mm. This medium is non-contaminating hence most suitable for immersion optical lithography. By immersion technology, DUV 157nm will be able to go up to minimum line width of less than 30nm.

** The mask and the refractive lens used in conventional lithography systems would absorb 100nm light required for patterning features smaller than 50 nm. Hence refractive lens is replaced by reflective mirror and masks are replaced by reflecting type Molybednum-Silicon masks a few nm’s thick . On this a pattern of chromium is laid down. This absorbs light in appropriate places and reflects elsewhere and thus the pattern is transferred to the Wafer.

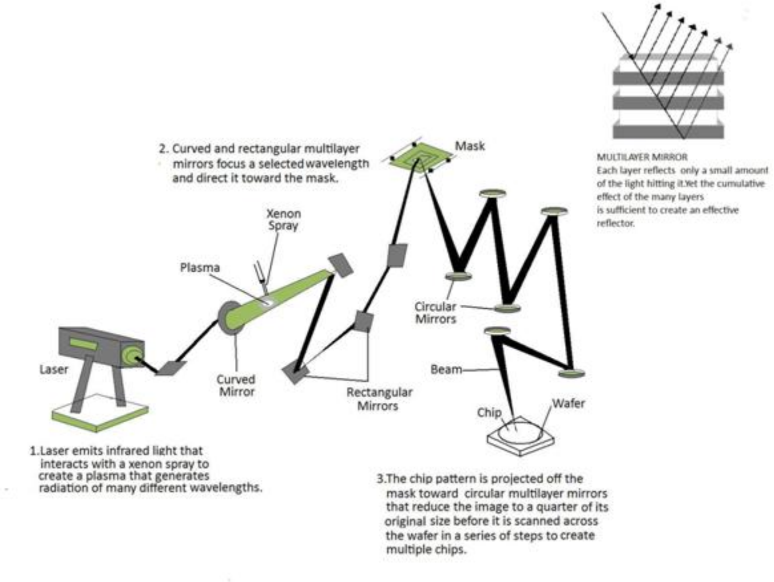

MULTILAYER MIRROR
Each layer reflects only a small amount of the light hitting it.Yet the cumulative effect of the many layers is sufficient to create an effective reflector.
2. Curved and rectangular multilayer mirrors focus a selected wavelength and direct it toward the mask.
Mask
Xenon Spray
Plasma
Circular Mirrors
Laser
Curved Mirror
Beam
Rectangular Mirrors
Chip
Wafer
1.Laser emits infrared light that interacts with a xenon spray to create a plasma that generates radiation of many different wavelengths.
3.The chip pattern is projected off the mask toward circular multilayer mirrors that reduce the image to a quarter of its original size before it is scanned across the wafer in a series of steps to create multiple chips.

Figure 3. Extreme UltraViolet Lithography.

[Curtsey *Scientific American* **, April 2004, The First Nanochips by G.Dan Hutcheson]**

At 100nm and below Optical Lithography reaches its limit even with Resolution Enhancement Techniques such as Optical Proximity Corrections and Phase Shift Masks. So we have to work out Post Optical Lithgraphy or Next Generation Lithography. SAMATECH Program has come up with five alternatives for NGL. These are described in Appendix II.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 4.

Summary: Part 4 of 50 years journey;

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 4.

Title:Part 4 of 50 yrs journey.

Keywords:Band-gap narrowing,BiCMOS,PolySi Emitter,Surface Recombination Velocity, Metal Silicide, base spreading resistance;

Summary: Part 4 of 50 years journey;

Section IV. The development in Bipolar Junction Transistor Structure over last 50 years for meeting the challenges of Post Industrial Society.

("Transistor Design and Application Considerations for $\geq$200GHz SiGe HBT", Grey Freeman et al, IEEE Transactions on ED, Vol 50,No.3, March 2003, pp.645-655)

BJT has historically been faster than MOSFET. FET depends on leading edge lithographic dimensions(45nm) but BJT depends on vertical base widths which has reached 10nm. BJT has structural flexibilities in minimizing the parasitic. It has higher transconductance, high self gain, low 1/f flicker noise and better V_{BE} matching. All these factors make BJT the device of choice for demanding applications. Today SiGe HBT are surpassing GaAs HBT . SiGe depends on more conventional Si Technology whereas GaAs depends on Molecular Beam Epitaxy which has a much higher overhead cost. So combination of HBT and CMOS, known as BiCMOS, with a large menu of passive devices such as R,C,L, Varactor Diode, and low loss transmission lines opens numerous possibilities and cost effective ways of realizing systems which cannot be achieved by GaAs ICs. BiCMOS reduces the cost of mobile consumer products, advance high BW wireless communication and collision-avoidance automobile radar. BiCMOS lead to VLSI, ASIC & Si based RF SOC solution. Hence SiGe/SiGeC is becoming the technology of future.

Growth of Personal Computers kept CMOS in limelight and fueled the CMOS scaling in 70s. Today the growth of INTERNET, high speed communication that converges data, video and audio and the growth of mobile communication has sharply brought in focus the need for cheap, high speed and low power devices. SiGe HBT is a major candidate for the needs of New Communication. DSP and DSP & Real World Interface are at the heart of INTERNET, so developing System on Chip (SOC) and integration of digital and Analog/RF function became the high priority area of research. Since advanced SiGe Process shrinks features and boosts performance it became the technology of choice for implementing 24GHz radar for blindside detection, for 77GHz radar system for automobile collision warning or advance cruise control, for 60Ghz Wi-Fi chips for next generation wireless LAN and backbone networks , for software defined radios, for cellular handset and for high frequency automatic test equipments.

The emergence of high performance SiGe Technology tailored for low voltage, low power RF & mixed signal applications has shown great potentials for microwave and mm wave applications.

IV.1. The relation between the physical parameters and performance parameters and Physics of High Performance nano BJT.

The carrier concentration profile in a forward biased diode is shown in Appendix III. As discussed in the Appendix III, in olden times EB junction was like a wide base diode but with dimension scaling and improvement in technology, EB junction became shallow and hence narrow base diode. In a narrow base diode the emitter surface contact is decisive in determining the resultant emitter injection efficiency. As shown in Figure 4, metal contact directly to pure single crystal emitter portion causes a poorer injection efficiency as compared to that where metal contact and n+ pure Silicon has heavily doped poly-Silicon emitter sandwiched between the two layers.

IV.1.1. Band Gap Narrowing in Emitter, its deleterious effects and its cure.

Heavy doping of emitter causes Band Gap Narrowing in Emitter which limits the improvement in Injection Efficiency due to heavy doping of Emitter. Emitter doping may be $N_D = 10^{20}$/cc but in effect we get the Injection Efficiency corresponding to 10^{18}/cc [Appendix IV]. Hence we say that BGN limits $N_{Deff} = 10^{18}$/cc. This effect(BGN) is mitigated by using heavily implanted PolySi layer used as emitter contact.

Due to heavy doping in Emitter, degeneracy is introduced which leads to Band Gap Narrowing(BGN) by ΔE_g in

$$\frac{(n_i^2)eff}{1} = \frac{N_C N_V Exp(\frac{-(E_{gE} - \Delta E_g)}{kT})}{1}$$

As shown in Appendix IV,

$$N_{Deff} = N_D Exp(-\frac{\Delta E_g}{kT})$$

For a doping level of $N_D = 10^{20}$/cc, $\Delta E_g = 0.12$eV.

This gives an effective doping level of $N_{Deff} = 10^{18}$/cc.

The classical formula for current gain without BGN is:

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{N_{D|E}}{N_{A|B}}\right)$$

If we take the following data: $(N_D)_E = 10^{20}$/cc, $(N_A)_B = 10^{17}$/cc, $W_E = W_B = 1\mu m$ and $D_{pE} = 1.25$ (cm)^2/sec and $D_{nB} = 20$ (cm)^2/sec then $\beta_F = 1.6 \times 10^4$;

But taking into account of BGN, $\beta_F = 160$.

The other component responsible for improved Current Gain in Poly_Si Emitter Transistor is due to flatter slope of excess minority carrier in Emitter region [Appendix V]

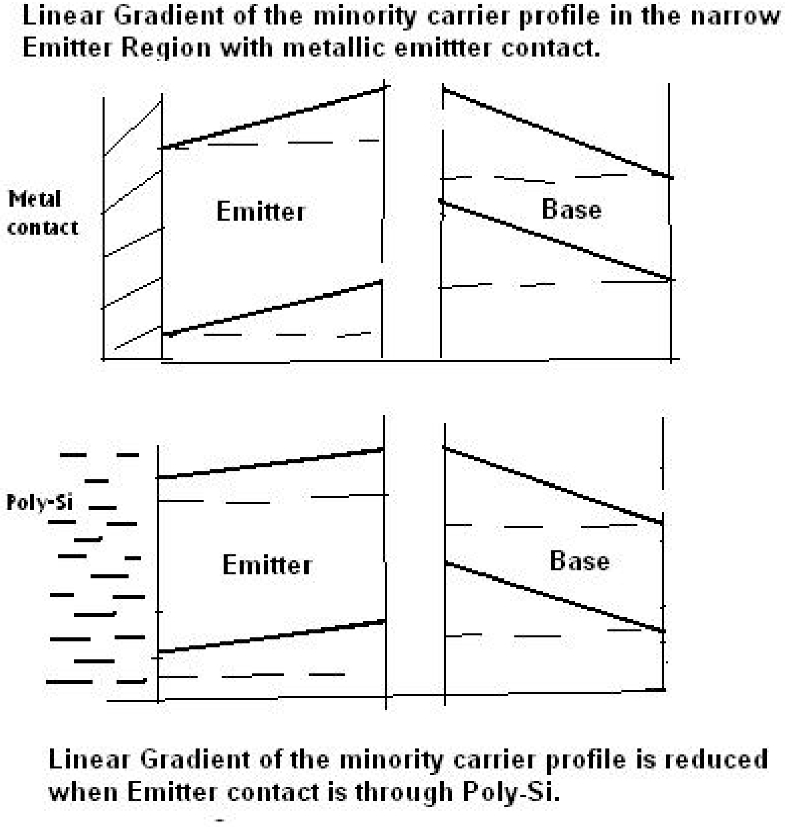

Linear Gradient of the minority carrier profile in the narrow Emitter Region with metallic emittter contact.
Metal contact
Emitter
Base
Poly-Si
Emitter
Base
Linear Gradient of the minority carrier profile is reduced when Emitter contact is through Poly-Si.

Figure 4. Linear Gradient is a strong function of surface condition. Metal contact gives a large gradient and Poly Silicon contact to Emitter gives a much lower linear gradient on the emitter side thereby improving I Dn /I D (Injection Efficiency) by several orders of magnitude.

To achieve higher degree of integration we had to go for smaller feature size as well as shallow devices. Metallic contact gives infinite surface recombination velocity, hence it gives a much steeper gradient as shown in Fig 4a. whereas Poly-Silicon gives a much lower gradient resulting in a very low hole component of the total current thereby giving a much higher injection efficiency. [Appendix V]

Table 7. Room Temperature current gain as a function of emitter contact for device run BIP-8.

[“Effect of Emitter Contact on Current Gain of Silicon Bipolar Devices”, T.H.Ning & R.D. Isaac, *IEEE Int. Electron Devices Meeting,* **pp.473-476, 1979].**

Wafer	H	H	E
Contact	Al	Pd_2Si + Al	Poly + Al
R_{DB} (KΩ/□)	7.2 ± 0.2	7.2 ± 0.2	7.6 ± 1.0
Gain	67 ± 3	51 ± 8	158 ± 12

“ The current gain of silicon bipolar transistors with shallow emitters depends critically on the emitter contact technology……………The conventional contact by metal or metal silicide degrades the current gain, while contact by a thin layer of polysilicon is effective in improving the current gain” [ibid].

IV.1.2. Base Spreading Resistance r x and the frequency response of BJT:

This is due to the narrow sandwich of high sheet resistance P base. P Base is of 200Ω/□ to keep $N_{A|B}/N_{D|E} = 10^{17}/10^{19}$ as low as possible. This results in $r_x = 100\Omega$. This creates serious deterioration in high frequency performance of BJT especially in terms of Unity Power Gain Frequency (f_{max}).

It can be shown that $(f_{max}) = [f_T/(8\pi r_x C_\mu)]^{1/2}$

Using the dimensions of BJT used in 70s we get [Appendix IV], we get

f_t (Unity gain BW)= 624MHz;

$f_{max} = [f_T/(8\pi r_x C_\mu)]^{1/2} = 18.6\text{GHz}$;

From this example it is clear that base spreading resistance r_x must be minimized to get best frequency performance but reduction of r_x requires increased base conductivity but increased base conductivity means lower Injection Efficiency because

Injection Efficiency $=\gamma = 1/[1+\sigma_B W/(\sigma_E L_p)]$

From injection efficiency point of view: $\sigma_B << \sigma_E$;

These are contradictory requirements hence to achieve Short Circuit Current Gain of 100 we cannot allow Base Spreading Resistance to go below 100ohm.

By using Poly Silicon , γ can be improved several orders of magnitude. This results in β of the order of 10,000. Here base conductivity can be increased to minimize Base Spreading Resistance. Thus with a reasonable β of 100, considerably lower Base Spreading Resistance can be achieved which plays crucial role in frequency performance and in improving the Figure of Merit of BJT which happens to be

GBP$= 1/(r_x C_\mu)$;

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 5.

Summary: This part 5 of 50 yrs journey.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 5.

Keywords: Bipolar Hetero structure, homo junction, energy band diagram, emitter injection efficiency;

Summary: This part 5 of 50 yrs journey.

IV.1.3. Method of improving Emitter Injection Efficiency by Bipolar Hetero Structure.

$I_C/I_B = \beta_F = [-qAD_n.\ n_p(0)|_B/W_B]/[\ -qAD_p.\ p_n(0)|_E/W_E] = (W_E/W_B)(D_{nB}/D_{pE})(n_p(0)|_B/p_n(0)|_E)$;

But

$$n_p(0)|_B = n_p^{\square} Exp\left(\frac{V_{EB}}{V_T}\right) = \frac{n_i^2}{N_A} = \frac{N_C N_V Exp(\frac{-E_{gB}}{kT})}{N_A}$$

$$p_n(0)|_E = p_n^{\square} Exp\left(\frac{V_{EB}}{V_T}\right) = \frac{n_i^2}{N_D} = \frac{N_C N_V Exp(\frac{-E_{gE}}{kT})}{N_D}$$

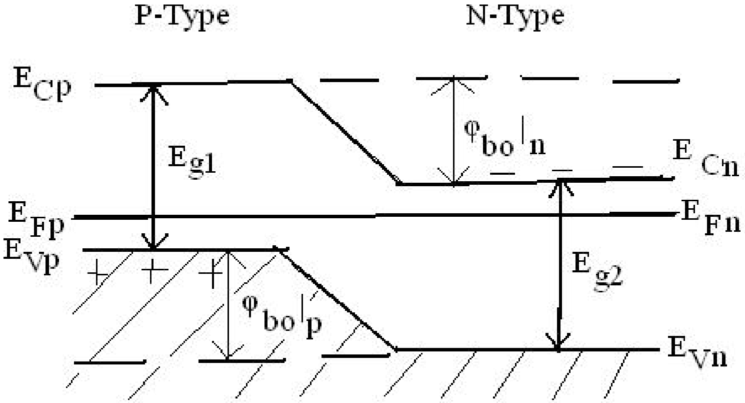
$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{N_{D|E}}{N_{A|B}}\right) Exp[\frac{(E_{gE} - E_{gB})}{kT}]$$

So if the EB junction is not a homojunction but a hetro junction

where $(E_g|_E - E_g|_B) = 0.15$eV then $\beta_F = 400$ even if $(N_D|_E/N_A|_E) = 1$ that is both E and B are equally doped. But $(N_D|_E/N_A|_E) = 1$ will drastically reduce Base Spreading Resistance which will leads to improved frequency response as well as improved GBP.

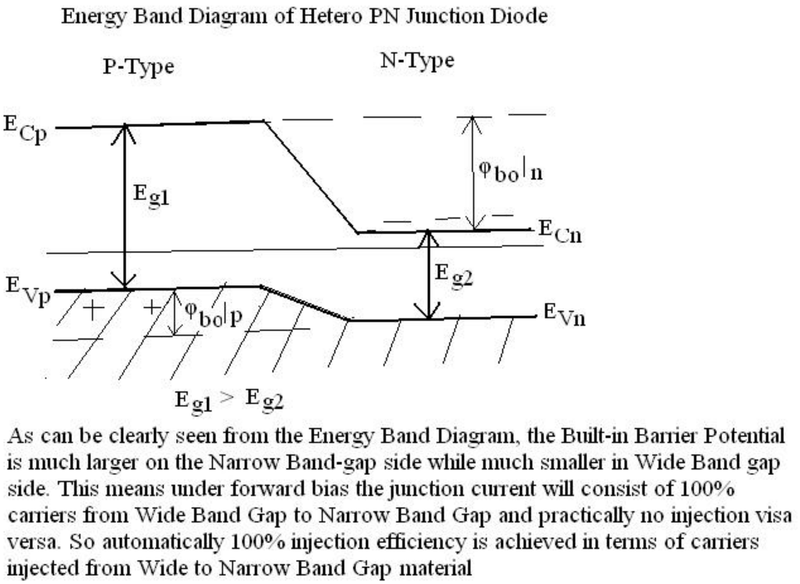
Graphically the improvement is evident as shown in the Graph:

Energy Band Diagram of a HOMO PN JUNCTION.



As can be seen in a HOMO JUNCTION, the built-in barrier potentials (are of equal magnitude on the two sides of the junction hence it is equally difficult for the minority carrier to cross over the potential barrier at the junction under forward bias. $E_{g1} = E_{g2}$

Figure 5. In a HOMO JUNCTION of equal doping on two sides the total forward junction current is made up of 50% electrons being injected from N to P side and 50% holes being injected from P side to N side.



As can be clearly seen from the Energy Band Diagram, the Built-in Barrier Potential is much larger on the Narrow Band-gap side while much smaller in Wide Band gap side. This means under forward bias the junction current will consist of 100% carriers from Wide Band Gap to Narrow Band Gap and practically no injection visa versa. So automatically 100% injection efficiency is achieved in terms of carriers injected from Wide to Narrow Band Gap material

Figure 6. In a HETERO JUNCTION under forward bias the Built-in Barrier Potential are so unequal that Junction Current is wholly constituted of carriers injected from Wide Band Gap to Narrow Band Gap thereby automatically achieving 100% Injection Efficiency even without creating large doping differential on the two sides.

Thus we see that heterojunction are very naturally disposed towards giving 100% injection efficiency even with a very low base spreading resistance.

SiGe Hetro Bipolar Transistors(HBT) hold the key to realizing high speed wire-line and wireless communication circuits and systems.[“Oveview: Fabrication of SiGe HBT & CMOS Technology” edited by Cressler,CRC Press, 2009]

In hetero Junction BJT we have two main contenders:

1. Si-SiGe-Si in IC Technology;
2. GaAlAs-GaAs-GaAs in MBE Technology.

History of SiGe HBT is given in Appendix VI. In this technology, pseudomorphic $Si_{(1-x)}G_x$ layer is used for the base. They are thin films less than 100nm that conform to the lattice pattern of the underlying substrate. This is strained crystalline structure. Carriers have high drift mobility under strained condition. If this film is grown any thicker than 100nm then the stressed film relaxes creating numerous defects and lattice mismatch which renders the film completely useless from device fabrication point of view.

SiGe HBT [Appendix VI] is the child of marriage of Si Technology and Bandgap Engineering Methods of III-V Compund Semiconductor. Only through low temperature Si epitaxy that the alloying of Si & Ge having different band-gaps is possible and doping of such layers with high precision is made possible. SiGe based bandgap engineering is Si Heterostructure Bipolar Technology (SiGe HBT).

SiGe can operate at Cryogenic temperatures whereas BJT freezes out at liquid nitrogen temperature. Hence superior performance parameters can be derived Cryognically cooled SiGe HBT.

SiGe HBT has complete tolerance towards radiation induced damages. Hence it finds very extensive space and interplanetary remote sensing applications. The advantage of low frequency noise of BJT is retained in SiGe HBT.

Though Carbon (C) contamination is very worrisome in minority carrier devices due to deep traps introduced by C, still 0.1 to 0.2% (i.e.10^{20}/ cm^{3}) proved to be very beneficial in Si:SiGe:Si devices as it suppressed out diffusion of Base Boron in subsequent thermal cycles.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 6.

Summary: This Part 6 of 50 years journey.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 6.

Keywords: Vertical Scaling, Lateral Scaling, transit time, scatter limited velocity;

Summary: This Part 6 of 50 years journey.

IV.1.3.1. The scaling issues in SiGe HBTs.

SiGe successful demonstration was first done in late 1980[Appendix VI,history of HBT]. This was continuously scaled to achieve a record peak f_T of 300GHz in 2006[“Effect of collector lateral scaling on performance of high-speed SiGe HBTs with $f_T > 300$GHz”, Rieh, Khater, Joseph, Freeman and Ahlgren, IEEE ED Letters, Vol 42, ,No20, 28th September 2006.]

The speed of operation in vertical devices has two components: transit time across the base and BC reverse biased junction on one hand and RC delays due to junction parasitic and bulk resistances.

$$\frac{1}{2\pi f_T} = \tau_{EC} = \tau_E + \tau_C + \tau_B + \tau_{CSCL}$$

$$= \frac{kT}{qI_C} C_{EB} + \left(\frac{kT}{qI_C} + R_C + R_E\right) C_{CB} + \frac{W_B^2}{YD_n} + \frac{W_{CSCL}}{2v_{sat}}$$

τ_E = RC delay time constant at EB junction;

τ_C = RC delay time constant at CB junction;

τ_B = Transit time delay through Base Width;

τ_{CSCL} = Transit time delay through reverse biased CB junction;

v_S = scatter limited velocity of electrons while falling down the potential hill at CB junction;

The vertical scaling is done for reduction in base and collector transit time. This is achieved by reducing the thermal cycles once base profile is achieved. Ge ramp in SiGe base has to be increased and C added to base for limiting the boron diffusion. Ge ramp creates a built-in field which aids the diffusing electrons cross the base region.

To reduce the time delays, base width and collector widths have to be reduced and I_C must be increased which requires Kirk Effect should be pushed to higher Current Desities. This achieved by increasing Collector Dopent Concentration. Hence Selectively

Implanted Collector(SIC) is introduced.

Lateral scaling reduces the overlap and cross sectional area of the junction hence parasitic capacitance and resistance will be reduced leading to a diminished RC delay time constant. Therefore along with aggressive vertical scaling a judicious lateral scaling also has to be resorted to. Excessive lateral scaling results in an early onset of Kirk effect. So lateral scaling has to be optimized.

Decreasing the Shallow Trench Isolation Width increases the perimeter collector to substrate capacitance which is a significant component in time-delay equation. Hence STI Width is not decreased every generation.

Decreasing the distance between SIC(selectively implanted collector) and extrinsic base implant will increase the overlap i.e. CCBOL. This has to be minimized. This is achieved by raised extrinsic base structure.

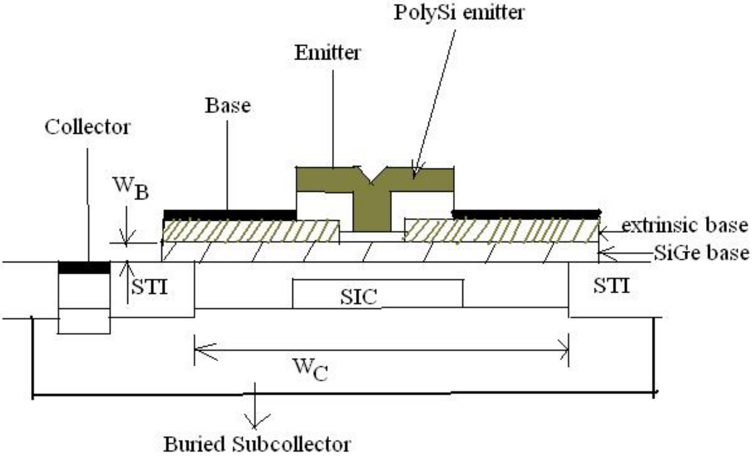


Figure 7. Cross-sectional view of fabricated device.(W_B is the base width which is scaled under vertical scaling and W_C is the collector width which is scaled under lateral scaling).

The result of aggressive vertical scaling and judicious lateral scaling is shown in the Figure 8.

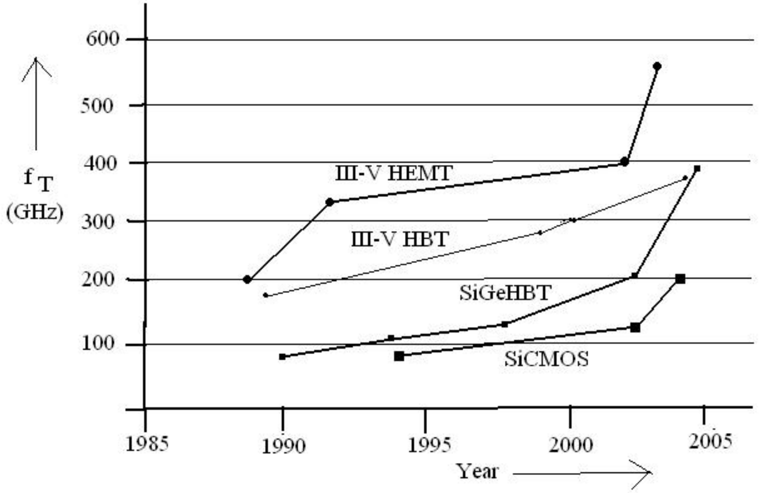


Figure 8. Cutoff frequencies versus year for various Technologies. [Rieh et al, IPRM, 2003, pp374-377]

As is evident from graph in Figure 8, today in 2009 through scaling and innovation , SiGe HBT is a better technology than III-V technology. While trying to improve the performance using scaling technology, some new problems come to the fore namely:

Self heating , electromigration and reduction in the safe region of operation. By scaling the emitter width(E_W) with respect to $1/J_C$ (the peak f_T current) significantly reduces self-heating and it eases electromigration problem[“ A look into the Future for SiGe HBTs”, Harame, Freeman et al,IEEE Symposium,WA3.5(invited)]

Heterojunction Bipolar Transistor strategy was adopted to maintain high injection efficiency with high Base conductivity so that base spreading resistance could be minimized and a better figure of merit as well as better f_{max} could be achieved but in practice it is found that as base doping is increased and Ge content and ramp increased, neutral base recombination is enhanced leading to a deterioration in current gain. So an optimization in Ge profile has to be done inorder to get the full advantage of increased base doping[“Current Gain of SiGe HBT under High Base Doping Concentrations” Ningyue Jiang & Zhenquieng Ma,” Semiconductor Science & Technology, 22, S168-S172, 2007.].

SiGe HBT technology was originally developed for high-end computing but it failed because it consumed much more power as compared to that in CMOS. Hence it was abandoned. Subsequently it was found that for RF circuits it consumed much less power as compared to that of CMOS for the same level of performance. So SiGe rapidly developed its application area in various forms of communications such as wired and wireless communication, disk storages, high speed high bandwidth instrumentation. Discrete SiGe are being used universally in RF amplifies and integrated SiGe chips are prolifigating in GSM,CDMA wireless hand sets and base stations. They are rapidly penetrating LAN chipsets and high speed 10-40 Gb/s synchronous optical network(SONET) transceivers.

SiGe real strength lies in its ability to integrate analog, RF and digital on a single chip using existing CMOS fabs. This kind of compatibility is not found with GaAs technology. Apart from this versatility it is enabling new architecture such as direct conversion and software radio.

Today SiGe BiCMOS is the fastest growing semiconductor process.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 7.

Summary: This Part 7 of 50 years journey;

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 7.

Keywords: Sidewall Oxidation Technique, Trench Isolation Technique, Reactive Ion Etching.

Summary: This Part 7 of 50 years journey;

IV.1.5. Sidewall oxidation and Silicon-on-Insulator for reducing the parasitic capacitances and improving the frequency response.

Large junction parasitic capacitances are introduced due to Silicon P-type substrate and due to sidewalls of isolation diffusion. Substrate capacitances can be eliminated by using Silicon-on Sapphire or by Silicon-on-Insulator. Isolation diffusion side-wall junction can be eliminated by using sidewall oxidation.

These structures are shown in Figure 7 and Figure 8. P+ isolation diffusion can be replaced by Trench isolation by Reactive Ions Etching (RIE) as shown in Figure 9.

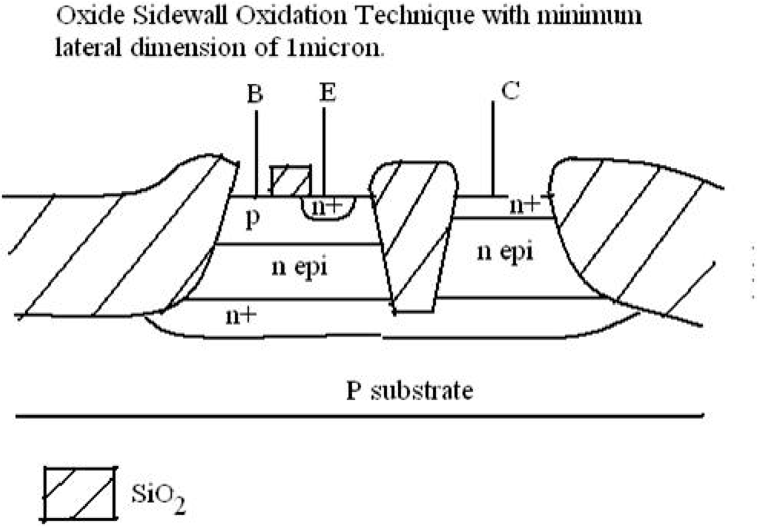
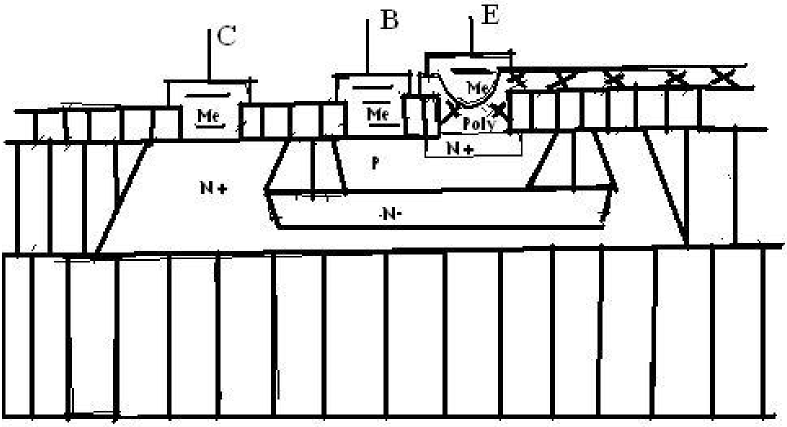

Oxide Sidewall Oxidation Technique with minimum lateral dimension of 1micron.
B
E
C
p
n+
n+
n epi
n epi
n+
P substrate
SiO2

Figure 9. Oxide Sidewall Oxidation Technique to minimize the parasitic capacitances.



Silicon on insulator with sidewall oxidation to reduce Junction parasitic Capacitance. Polysilicon Emitter Contact acts as a source for shallow emitter diffusion with self alignment automatically occurring thereby reducing the effective cross sectional area of EB junction.

▥ SiO$_2$ ☰ Metal ☒ PolySi(ion implanted)

Figure 10. Cross sectional view of an advanced self aligned polysilicon Emitter BJT with side wall oxidation and Silicon on Insulator for reducing the parasitic capacitances.

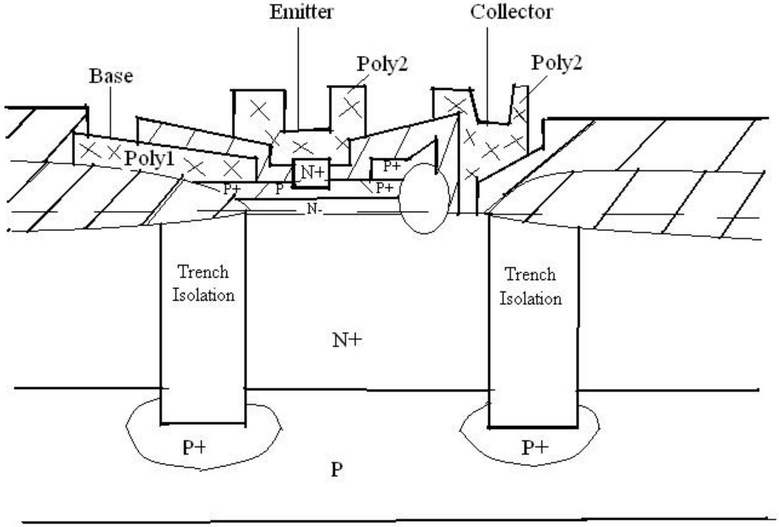


Figure 11.Cross-Section diagram of an advanced double self-aligned polysilicon emitter BJT with sidewall oxide and trench isolation.

[10.7GHz BJT using double layer process technology, Electroniuc Letters Vol.24,920-22, July 1988]

P+ isolation diffusion is replaced by trench isolation. Trench Isolation is achieved by Reactive Ion Etching(RIE). Emitter and Base regions are self aligned to minimize AE. Oxide side wall used for reducing junction capacitance.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 8.

Summary: This is the concluding part 8 of 50 years journey.

The Journey of I.C.Technology from micro (1959) to nano (2009) era.-Part 8.

Title: Part 8 and the concluding part of 50 years journey.

Keywords: High K- Metal Gate(HK-MG), Atomic layer deposition, Fermi-Level Pinning(FLP),Strained Silicon,SiGe:C HBT,

Summary: This is the concluding part 8 of 50 years journey.

IV.1.6. High-k solution for ULSI CMOS. [The high-k solution by Mark T. Bohr, Robert S. Chau, Tahir Ghani & Kaizad Mistry, *IEEE Spectrum* , October 2007, pp 23-29]

As the level of integration evolved, size of CMOS halved every 24 months. At this rate of scaling, by 1998 SiO_2 gate insulation became 5 atomic layer thick with total thickness scaled to 13Aº each atomic layer being 2.6Aº. Current leakage and heating became a serious problem. Because of wave nature of electron and its quantum mechanical tunneling property, thin gate oxides allow the electrons accumulated on gate to leak to the channel. To overcome this problem we need physically thick gate oxides to prevent quantum mechanical tunneling but at the same time electrically thin so that channel is turned at low threshold. If dielectric constant is doubled then thickness can be doubled without any reduction in Turn-ON capability since $C=k\varepsilon_0 A/d$

$$C = \frac{k\varepsilon_0 A}{d}$$

where k = dielectric constant or relative permittivity of the gate insulator

d = thickness of the gate insulator and A = cross-sectional area of the gate insulator.

Metal interconnects are preferred over low-k material so that propagation delay is minimized.

$$v_{propagation} = \frac{1}{\sqrt{\mu_0 k \varepsilon_0}}$$

Low k will give higher velocity of propagation thereby minimizing the propagation delay.

MOS capacitance had to be reproducible, its behavior should be repeatable. To achieve this reproducibility and repeatability, Reactive Sputtering or Metal Organic CVD had to be abandoned in favour of Atomic Layer Deposition (Appendix III). This gave the necessary smoothness to the surface of the electrode as well as precise control over the thickness of the Gate Insulator. This ensured reproducibility and repeatability.

Several dielectrics were studied such as Al_2O_2 , TiO_2 , Ti_2O_2 , Ti_2O_5 , HfO_2 , $HfSiO_4$, ZrO_2, $ZrSiO_2$, La_2O_3 .

With scaling due to Fermi-Level pinning (AppendixIV), higher threshold Voltage (V_{Thres}) resulted. Also High-k material have high elasticity hence result in higher phonon scattering or lattice scattering resulting in lower channel mobility. By screening the phonon effect, the deterioration in channel conductivity ould be prevented. This required by increasing the electron density in Poly-Si gate. Hence we had to switch to Metal-High k combination. This prevented:

1. Fermi-Level pinning hence threshold voltage normalized;
2. This normalized channel conductivity which had drastically deteriorated due to dipole vibration effect;
3. Metal Gate Electrode provides better bonding between Metal-Dielectric. Hence table operation.

In mid-2003, Intel's Hillsbaro, Ore, Development Fab developed HK-MG CMOS. Intel's 130nm

technology was used. Using Hafnium-Based Oxide and Metal Gate electrode following characteristics were achieved:

1. Low Threshold = 1V;
2. Negligible leakage current through Oxide;
3. High 2-D channel mobility;

The standard fabrication method was "Gate First". In this method:

Gate Dielectric + Gate Electrode were laid first;

Source and Drain implanted;

Silicon is annealed to repair the damages that occurred during ion-implantation. High Temperature

became a problem for the new technology of HK-MG. So "Gate Last" technology was adopted which circumvented the annealing problem. This led to a paradigm shift in late 2004.

The new flow process was 45nm technology with High k + Metal Gate using Gate-Last strategy. Using this flow process in January 2006, 153Mb SRAM with 1 billion CMOS was built. Leakage gate current was reduced by a factor of 10. But there was sub-threshold leakage.

Scaling had reduced Threshold Voltage but reduced V_{Th} led to increased sub-threshold leakage which defeated the nanoWatt-logic objective. Each new generation of scaled down transistor would increase I_{ON} by about 30% but would lead to $I_{sub\text{-}threshold}$ increase by 3-Fold. So at ULSI level, power efficiency and low leakage would be at premium rather than speed.

Table 9.The options at 45nm HK-MG CMOS Technology

	Option 1	Option 2
Oxide thickness	13Aº	26Aº
I_{ON}	25% increase	No increase
$I_{sub\text{-}threshold}$	No increase	1/5 I_{sub}
V_{Thres}	Same as before	Increased threshold

The CMOS circuits are built between these two extremes.

PENRYN dual-core μP has 410mCMOS and PENRYN quad-core μP has 820 m CMOS. These are optimized for mobile applications, desk-top computers, 64-bit workstations and server applications.

IV.1.7.Mobility enhancement in strained Silicon.

[Mobility Enhancement, the next vector to extend Moore's Law, Nidhi Mohata & Scott E. Thompson, IEEE Circuits & Devices Magazine, September/ October 2005, pp.18-23]

Geometric Scaling has been driving IC Industry till date. Since 90-nm Technology generation was introduced, off-state leakage current and power density have made scaling a difficult and challenging job. New scaling vectors were adopted to meet this challenge. At 90-nm generation, mobility enhancement through uniaxial process-induced strained Si has emerged as the next scaling vector.

The theoretical formulations of carriers in 2-D inversion layer just below the insulator Gate:

$$\mu(carrier\ mobility) = \frac{q\tau}{m^*}$$

where τ is the mean free time between two consecutive scatterings;

$1/\tau$ = scattering rate;

m^* is the conducting electron effective mass.

Under strain, both m^* and scattering rate reduce leading to enhanced mobility.

Uniaxial stress always provide higher current enhancement(1.46mA/µm and 0.88mA/µm for n-channel and p-channel respectively) as compared to that produced by biaxial stress(0.85mA/µm and 0.45mA/µm for n-channel and p-channel respectively). Hence in state of art technologies, Industries have adopted uniaxial stress. There are three state-of-art techniques:

1. A local epitaxial film is grown in the source and drain regions;
2. In the second technique we use a capping layer;
3. A dual capping approach.

FUTURE DIRECTIONS OF IC INDUSTRIES.

In the future we hope to go from 45nm technology to 32nm technology to 22nm technology to 16nm technology. The demand for high-speed transistors is increasing with the band-width required for telecommunication systems. Combining aggressive vertical scaling with reduced device parasitic, it has been possible to achieve 210GHz at room temperature using SiGe:C HBT Technology [“ A 210-GHz f_T SiGe HBT with a non-self-aligned structure”, Jeng, Jagannathan et al, IEEE ED Letters, Vol.22, No.11, Nov 2001] . Safe operating voltage measured is reasonable.

Table 10 . Measured values of Safe Operating Voltages.

Parameter	Measured Voltage(V)
BV_{EBO} (V)	3.3
BV_{CBO} (V)	5.5
BV_{CEO} (V)	1.8
Dc current gain	450

This has established that SiGe HBT Technologies can be extended to high frequency and high bit rate applications.

SiGe HBT will continue to scale in future generations. Recent research demonstrations have achieved npn SiGe HBTs with peak f_T of 350-375GHz[Rieh,Jaganatthan et al “SiGe HBTs with cut-off frequency of 350GHz,” IEDM Tech.Digest,2002, 771-77] and 380GHz [Heinemann, Barth et al, “Low parasitic collector construction for high-speed SiGe:C HBTs, IEDM Tech.Digest,2004, pp.251-254] . Through simulation studies it has been suggested that through suitable lateral and vertical scaling , SiGe HBT of f_T above

500GHz is possible on 100% silicon(CMOS) fabrication compatible [“Half-Terahertz Operation of SiGe HBTs” Kritivasan, Cressler, Rieh,Khater, Ahlgren and Freeman, EDLetters, Vol 27,No.7, July 2006, 567- 569].

Table 11. Transistor Performance Summary at 300, 77 and 4.5K for an emitter geometry of $0.12 \times 1.0 \mu m^2$.

Prameter	300K	77K	4.5K
Peak β	1499	9297	12554
Peak f_T (GHz)	352	456	510
Peak f_{max} (GHz)	241	295	276
r_x(base spreading resistance) at peak f_T (Ω)	304	206	305
Transit Time (psec)	0.36	0.33	0.27
J_C @ peak f_T (mA/μm^2)	27	24	21
BV_{CEO}(V)	1.47	1.43	1.36

Today SiGe technology has become the driving technology behind the Information Revolution. SiGe is enabling lightweight, personal communication devices like digital wireless handsets and other entertainment and information technology devices such as digital set-top boxes, Direct Braodcast Satellite, automobile collision avoidance systems and personal digital assistants.

SiGe extends the life of wireless phone batteries. It allows smaller and more durable communication devices. SiGe technology is being used for developing mobile-computing systems which converge the capabilities of cellular phones, global positioning systems and internet on one platform. These SiGe based multifunction, low-cost, mobile client devices are the building blocks of future computing. These devices will communicate over voice and data networks.

SiGe real strength lies in its ability to integrate analog, RF and digital on a single chip using existing CMOS fabs. This kind of compatibility is not found with GaAs technology. Apart from this compatibility resulting into versatility, it is enabling new architecture such as direct conversion and software radio.

Today SiGe BiCMOS is the fastest growing semiconductor process. There are 25+ SiGe Industrial Fabrication facilities all over the world.

Every major player in the field of Semiconductor Technology is either having in-house R&D facility in SiGe HBT Technology or it is having its group of designers use some body else’s facilities. The prominent names today in SiGe HBT industrial production are: IBM, Jazz Semiconductors, Hitachi, Infineon, IHP, ST Microelectronics, Texas Instruments and Philips.

The success of this technology is the compatibility of two promising technologies without any compromise or loss of any of its individual good qualities.

CMOS seems to have road blocks beyond 45nm node as is evident from the table below.

Table 12. HP NMOS Scaling Scenario.[“MOSFET Performance Scaling-Part II: Future Directions”, Ali Khakifirooz and Dimtri A. Antoniadis, *IEEE Transactions on ED*, Vol.55, No.6, June 2008,pp.1401-1408.]

Gen. (nm)	250 Poly/Si	180 do	130 do	90 do	65 do	45 (MG-HK)	32 do	22 do	15 do
Contacted Gate Length (nm)	910	636	445	310	220	155	110	78	55

Pitch									
Physical Gate Length	L_g (nm)	180	130	65	45	35	30	26	22	15
Physical Oxide Thickness	$T_{ox,phy}$ (nm)	4.5	2.5	2	1.2	1.3	4	3.4	2.5	1.2
Supply Voltage	V_{DD} (V)	1.8	1.5	1.4	1.2	1.2	1.0	0.9	0.8	0.7
Off Current	I_{off} / W (nA/μm)	0.5	3	10	40	100	200	300	300	300
Silicide Contact Resistance	ρ_c (10^{-8} Ω.cm^2)	9	8	6	5	4	3	2.5	2	2
Sheet resistance	R_{sheet} (Ω/sq)	300	250	200	200	200	200	200	200	200
Virtual Source Velocity	v_{xo} (10^6 cm/s)	8.1	8.6	9.2	10.8	13./8	16.5	16.5	16.5	16.5
Sat. Threshold Voltage	V_T (V)	0.51	0.45	0.42	0.36	0.36	0.33	0.31	0.31	0.31
Sat. Drain Current	I_{Dsat} (μA/μm)	694	937	1103	1362	1652	1761	1620	1450	1351
Source Series Resistance	R_s (Ω·μm)	97	93	84	86	85	87	93	103	114
Effective Fringing	$C_{f,eff}$ (fF/μm)	0.60	0.71	0.61	0.74	0.78	0.74	0.85	1.04	1.01

Capacitance									
Transistor Delay (ps)	6.26	4.56	2.55	1.88	1.46	1.15	1.23	1.47	1.45

This Table 12 is based on projection and simulation for nodes at 45nm and below. For nodes above 45nm, the Table is based on actual experimental work. It is clear that at 45nm and at greater scaling, we get counterscaling effects. Because of increased fringing capacitance, transit delay deteriorates beyond 45nm technology. But with emergence of multicore processors, number of available transistors is more important than making them faster. Hence scaling will continue to be used even though it is becoming counter productive.

The Present Status of Industrial Production of IC Chips.

There are two distinct IC manufacturers:

1. Chip Foundaries;
2. Dedicated Manufactureres;

In the past dedicated manufacturers designed and fabricated IC chips according to their own needs. But as the cost of fabrication soared up, dedicated manufacturers became “fab-lite”. In Fab-lites initial design is done by the manufacturer but the fabrication part is done in the Chip Foundary. AMD, Freescale and Texas Instruments have all turned fab-lites.

The major Chip Foundaries are :

1. Taiwan Semiconductor Manufacturing Co., Hsinchu, Taiwan;
2. United Microelectronics Corp., Hsinchu, Taiwan;
3. Chartered Semiconductor, Singapore;
4. Semiconductor Manufacturing International Corporation, Sanghai, China;
5. IBM Microelectronics (dominant foundary of Games Processors), EastFishkill, New York, USA;
6. Global Goundaries, a joint venture of AMD & Abu Dhabi.

1. 50% of the foundaries are utilizing a generation back methods of process technology which is 65nm node presently.
2. 40% use the older and larger technology;
3. Only 10% use state of art to compete with Intel and Samsung.

To this 10% belong companies which make graphics processors, FPGAs and moble phone chips. They have to continuously improve their design to include new features and they have to keep improving their speed of computing. This 10% have the requirement of next generation Lithography.

Conclusions.

We have come a long way since the first IC chip was invented in 1959 in TI and FairChild simultaneously. The invention of IC Chip has marked the beginning of a new historical era in Man-Kind’s history. This invention has been as crucial as the discovery of fire was or as momentous as the invention of wheel was.

Discovery of Fire brought the proto-human society down from its arboreal abode to its humble cave dwellings. This enabled the development of our society in rapid strides.

On the other hand the invention of wheel led to rapid exchange of ideas and commodities enabling a homogenization across the globe which would have been impossible otherwise.

In a similar fashion IC Technology has been a great leveler of human societies. All human beings are becoming a knowledge worker in one way or the other and the whole world is being rapidly transformed into a big Global Village where all new ideas or new technologies are available to one and all in the shortest time.

Part 9_Journey of IC

This last part of the main text of IC journey from 1959 to 2009 gives cost metrics of IC Chip. It gives the algorithm to determine the cost of a die.

Reference materials:ECE4121/ECE4141,VLSI Design,Design Metrics by AHM Zahirul Alam.

Cost of IC Chip = Non-recurring expenditure(NRE) + Recurring Expenditure(RE)

NRE is the fixed cost of mask generation. It is one-time cost factor.

RE is the expenditure incurred in Silicon procesing, packaging, volume of production and the die area.

70nm ASIC incurs $4M NRE on account of masks.

"The club of people who can offer an extreme sub-micron ASIC or COD design is getting pretty exclusive."Ron Wilson,EE Times (May,2000)

The Cost Metrics of the IC Chip.

Chip	M.L	L.W(µm)	W.C($)	W.D(mm)	D.S.	D(mm2)	D/W	Y(%)	D.C.

					cm2	(mm)				($)
386DX	2	0.9	900	150	1	10	43	360	71	4
486DX2	3	0.8	1200	150	1	12	81	181	54	12
P.PC601	4	0.8	1700	150	1.3		121	115	28	53
_7100	3	0.8	1300	150	1		196	66	27	73
DECα	3	0.7	1500	150	1.2		234	53	19	149
S.SPARC	3	0.7	1700	150	1.6		256	48	13	272
Pent.	3	0.8	1500	150	1.5	18	296	40	9	417

M.L.-metal layers, L.W.-line width,W.C.-Wafer Cost,W.D.-Wafer Diameter(mm),D./cm^2-defects per cm square,D.S.-die size(mm),D(mm^2)-Die Area(mm square),D/W-Dies/Wafer, Y-Yield,D.C.-Die Cost($)

Algorithm for calculating the die cost:

Recurring costs:

Variable Cost =

$$\frac{cost\ of\ die + cost\ of\ die\ test + cost\ of\ packagimg}{final\ test\ yield}$$

Where : Cost of Die =

$$\frac{cost\ of\ wafer}{\frac{dies}{wafer} \times die\ yield}$$

$$\frac{die}{Wafer} = \frac{\pi \times (\frac{wafer\ dia}{2})^2}{die\ area} - \frac{\pi \times wafer\ dia}{\sqrt{2} \times die\ area}$$

die yield= [1+((defect/area)*(die area))/alpha]^(-alpha)

One example:

Wafer size = 12” = 30.48 cm;

Die size = 2.5 cm2 ;

1 defect/cm2;

alpha=α= 3(complexity);

Number of dies per wafer = 252 dies /wafer;

Die yield = 16 %;

252×0.16 = 40 die/wafer;

40 die per wafer can actually be marketed.

Hence die cost = wafer cost/40 ;

Die cost is a strong function of die area.

Die cost is proportional to the Nth power of die area where n=3 to 4.

50 years Journey of IC Technology_Appendix I

This describes the history of development in the fields of Sold State Technology, Electronics,IC Technology, Computer, Super-Computer, microprocessor chips , local area network(LAN) and submarine cables.

Appendix I

Table I.1. Road Map of development of Solid State technology.

[“A brief history of semiconductors” by Tudor Jenkins, Physics Education, 40(5),pp.430-439;

“ The lost history of the Transistor- How 50 years ago, Texas Instruments and Bell Labs pushed electronics into the Silicon Age” by Michael Riordan, IEEE Spectrum, May 2004, pp.44-49]

Year	Invention
1731-1782	Gray(1731), Desa Guller (1742) and Volta (1782) investigate conduction in solids.
?	Humphrey Davey

	establishes the positive Temperature Coefficient of Resistance (t.c.r)of Metals.
1833	Michael Faraday detects the negative t.c.r. of AgS(Silver Sulphide). He also observes Thermal Runaway, contact potential at the interface of dissimilar materials and surface states.
1839	Alexandre- Edmond Becquerel observes Photo-Voltaic effect in electrolysis bath containing aqueous HNO_3 as electrolyte and AgCl coated Platinum electrodes.He also observed PV effect at matal-semiconductor junction in this case Ag/ AgCl interface.
1873	Witloughby Smith detected Light Dependent Resistance effect in Selenium while working on Submarine Cables but he identified it as metal. It was established as

	semiconductor in1907.
1876	William Grylls Adams and Richard Evans Day discovered PV effect in Selenium.
1883	PV Cell was invented by Charles Fritts. Selenium was deposited on metal substrate and Gold leaf film used as the second electrode. Cross-sectional area was 30cm^2 and thickness was only 30μm. It was working on only 1% efficiency. This device caused rectification also.
1874	Carl Ferdinand Braunn made a Point Contact Diode by pressing a wire point on galena(PbS). This was popularly known as Cat’s Whisker. This showed rectification. Similar rectification was observed in untarnished and tarnished Copper.
1887	Hertz had generated Radio Waves and it was detected by coherer. Soon Cat’s Whisker replaced the coherer.

1901	J.C. Bose got a patent for the point contact diode.
1906	Greenleaf Whitteir Pickard (an employee of AT&T) filed a patent for Si Crystal Detector for Radio Waves. The mechanism was not clear. It was hypothesized as thermoelectric effect or as electrolytic effect or as gas discharge.
1915	Carl Benedicks proposed Ge-Pt and Ge-Cu as point contact diode to be used as peak diode detector in Radios. But by then Vacuum Diode had been commercialized and it took the role of detection.
1920	CuO + Se was developed as PVC used in battery chargers, exposure meters in photography. It was also used as modulators and non-linear circuit elements.
1930	Alan F. Wilson published his seminal paper titled “Quantum Theory meets Semiconductors”. Band

	Theory of Solids, thermal generation of electron-hole pairs, bi-polar conduction, doping and doped semiconductors- all these phenomenon were theoretically understood. The requirement of pure semiconductor was urgently felt in order to fabricate stable, reproducible and reliable devices.
1940	Point Contact Diodes were used in RADARS at 300MHz & 3GHz.
1942	By melting, cooling and recrystallization and slicing the relevant portions 99.999% pure Electronic Grade pure Polycrystalline Silicon was obtained. This exhibited rectification as well as PV effect.
1942	Karl-Lark Horowitz at Purdue University stared work on PbS and then moved to the study & production of high purity Ge. Pure Ge gave high

	quality Cat's Whisker.
1946	Shockley, Bardeen & Braittain started workin on Si & Ge and on the equivalent of Triodes and Pentodes(tubes are short lived and high power consuming) at BELL Labs.
16th Dec 1947	Point Contact Ge Transistors were invented and power gain was observed.
24th Dec. 1947	Positive feed back was applied and sustained oscillations were observed. This was the litmus test for power gain.
30th June1948	The invention of Point Contact Ge transistors was publicly announced.
	Point Contact Transistor was difficult to make, difficult to control and unstable. So Bipolar Junction Transistor was proposed with N Type thin layer sandwitched between P type layers. Initially poly-crystal Ge was used to make BJT. These were

	irreproducible, unpredictable and low quality because of scattering from grain boundaries. Hence switch was made to single crystal Ge.
1948	Gordon Teal & John Little grew Single Crystal Ge using Jan Czochralski Method developed in 1918.
April 1950	At Bell Labs, Gordon Teal and fellow Chemist Morgan Sparks successfully made single crystal Ge BJT.
Feb 1951	Ge BJT has plus points such as less reactive, lower melting point and higher mobility hence higher frequency response. But it has serious negative points. These have higher leakage currents (due to low band gap energy of 0.67eV) which are very temperature sensitive almost doubling for every 10ºCelsius rise in

	temperature hence device becomes useless above 75ºC. Therefore Ge BJT could never be the device of choice for armed services which needed stable, reliable equipment that performed under extreme conditions. Hence switch was made to Silicon which could operate over much larger temperature range. Teal and Buehler grew pure single crystal Si and fabricated Si Junction Diode.Calvin Fuller was developing the diffusion techniques also at BELL Labs.
1952	John E Sabey at GEC fabricated Alloy Ge Transistor. This went for mass production.
Early 1954	Calvin Fuller & Gerald Pearson formed PN Junction diode by diffusion.
25th April 1954	BELL Labs announced the successful fabrication of large area Junction Diode

	which could act as Solar Cell with 10% efficiency
10 May 1954	At Institute of Radio Engineers National Conference on Airborne Electronics, Dayton, Ohio, Gordon Teal then at Texas Instrument announced the fabrication and performance of diffused Si BJT.
January 1954	Morris Tanenbaum at BELL Labs had also achieved a Si BJT by rate growing technique but they did not announce it.
The decade of 50s	TI had Si BJT market to itself for rest of the decade and thus the stage was set for the emergence of TI as the International giant in field of device manufacturing.
July 1954	Charles Lee made a diffused Ge BJT with few µm thin Base operating with f_T = 500MHz.Diffusion was controllable and could yield narrow bases.
17March 1955	Morris Tanenbaum

	fabricated a diffused Si BJT with few µm thin Base operating with f_T = 120MHz.
	Rest is history. We will see the birth and evolution of Integrated Circuits in the road map of Electronics.

Table I.2. Road Map of development of Electronics & IC Technology.(1876-1959&1959-2009)

[“One hundred years of Electronics” by Frank Thompson, Physics Education, 40(3), May 2005, pp252-256]

Year	Invention
1876	Alexandre Grahm Bell patents Telephone
1878	Thomas Alva Edison patents incandescent lamp.
1895	J.J Thomson discovers electron as the basic building block of Atom

	through the study of Cathode Ray.
1896	Marconi introduces the concept of Wireless Communication more commonly known as Radio Communication.
1904	John Ambrose Fleming invents & patents Vacuum Tube Diode(two electrodes device).
1906	Lee De Forest invented Vacuum Triode (three electrodes device). This becomes the work horse of Early Electronics. and used it to achieve RC coupled Amplifier.
1909	Marconi and Jagdish Chandra Bose simultaneously announce the invention of radio transmission and reception.
1909	Around the same time, point contact solid state diode or cat’s whisker was introduced as detector in radio Receiver.
1911	Using Heaviside “Distortion-less

	Transmission Condition” i.e. loading the Telephone cables with large inductors, first long distance telephone call was made from New York to Denver, Colorado, over a distance of 2000 miles.
1913	Vacuum Tube Repeater introduced in Public Service Telephone Network (PSTN).
1915	Using inductor loading & R.C Coupled Amplifier, first transcontinental telephone call is made from New York to San Francisco across a distance of 4000miles. Telephone becomes a Public Utility Service.
1919	Hull patents 4 electrode Tetrode. This had Cathode, Grid, Screen and Anode. Because of back emission of electrons from the anode , the output characteristics had a kink. So this device could not become popular.
1926	Jobst and Telegn

	introduced a fifth suppressor electrode to prevent back emission and remove the kink. This was known as Pentode.
1945	ENIAC (Electronic Numerical Integrator & Calculator) build by Moore School of Electrical Engineering at the University of Pennsylvania. ENIAC is 27Metric Tonnes , consumes 200kW and occupies a large room. The valves had a high failure rate of one per 10 minutes therefore valve based computers could only have niche applications. Valve based computer could not become a house hold item.
1945	First Transoceanic Telephone Links established using voice quality coaxial cables. Telegraphic Cables are not suitable for Telephone Conservation. Voice

	quality cables were developed during World War II and introduced right after the War .Before the World War II , no TransAtlantic or TransPacific Telephonic Links were there. Across the oceans only Radio-Telephone call could be made.
1947	At Bell Laboratories, John Bardeen, Walter Brattain & William Shockley invented BJT (Bipolar Junction Transistor). They were awarded Nobel Prize in 1953 in Physics for this invention.
1952	UNIVAC I (Universal Automatic Computer)- First computer designed for business use. The inventors were John Mauchy & John Eckart.7.26 Metric Tonnes, 5000 Vacuum tubes & 1000 calculators per second. These are termed **FIRST GENERATION**

	COMPUTERS.
1956	UNIVAC II using transistors come in the market. Manufactures are Sperry Rand, USA. These are termed as **SECOND GENRATION COMPUTERS**.

Road Map of development of Integrated Circuit Technology(1959-2009)

1959	**Integrated Circuit (IC)** chip invented by Jack Kilby (Texas Instrument) and Robert Noyce (FairChild). Discrete active and passive devices are integrated together on one wafer thin silicon chip in the area of 1cm by 1cm square, mounted on a header and then properly encapsulated and sealed. At the time it started, it had only 30 components in the 1 square cm area. This heralded the Information Age and **THIRD Information Revolution**. It also produced an

	Explosion in electronic consumer and communication applications such as Personal Computers to cell Phones. Jack Kilby was awarded the Nobel Prize in physics in the year 2000. Jack Kilby passed away on 1st August 2005. Robert Noyce went on to establish “ INTEL” (Integrated Electronics) along with Gordon Moore and Andy Grove . He passed away earlier than Jack Kilby so he could not be awarded. Noble Prizes are not awarded posthumously. Today a multiprocessor chip has billion transistors in the same chip size and can perform as well as a Supercomputer.
1961	FIRST RLT (Resistance Transistor logic) IC Chip is commercially marketed by T I (Texas Instrument) and FairChild.

1964	FIRST Analog IC Amp. Chip MC 1530 marketed by Motorola. Op. Amp. stands for Operational Amplifier which were in yesteryears used in Analo Computers for different Mathematical Operations and hence the name Op. Amp.
1964	PDP 8, table top computer introduced in the market .It was a 8-bit computer meaning by the address size is 8 bit. Larger is the address word, larger is the address capability. 4-bit means 16 locations can be addressed and 8-bit means 256 locations can be addressed. This used IC Chips hence it was compact and less power consuming. This was **THE THIRD GENERATION COMPUTER.** Price $18,000.00, 125 kg, and 4kBytes memory.
1968	18th July, INTEL (Integrated Electronics)

	was established by Robert Noyce and Gordon Moore . Later this team was joined by Andy Grove. Today 80% Personal Computers are equipped with INTEL Microprocessors.
1970	Advance Research Program Agency (ARPA) established ARPANET, the first WAN (Wide Area Network), connecting Stanford Research Institute, University of Utah, University of California Los Angles and University of California Santa Barbara mainly for Defense research purposes. In three years it grew to cover whole USA.
1970	PDP-11 introduced with 16 address word.
1971	**FIRST INTEL MICROPRPCESSOR CHIP (μP 4004)** introduced in the market . It can handle 4-bit data word and 8-bit address word meaning by it can

address 256 locations. This 1 chip combined the function of 12 subsystems of a CPU(Central Processor Unit). This one silicon chip (1cm × 1cm)had 2300 transistor , system clock rate 108kHz, Cost $299.00 . In India it was available for Rs.7500.00. [A development contract from a Japanese Company , for a set of chips needed to power an electronic calculator , triggered off what became the world's **first computer-on-a-chip: the Microprocessor.** 35 years ago, INTEL, a small Santa Clara (US) based manufacturer of memory and switching devices, unveiled a thumbnail-sized slab of silicon encased in a ceramic casing with 16 pins: this ushered in the era of Personal Computers, the era of Desk Top

Computing and the era of **FOURTH GENERATION COMPUTERS.** The contract with Busicom required INTEL to deliver the electronics for calculating machine, as a set of 12 custom built chips. The electrical engineer assigned to handle the task-Dr. Marcian “TED” Hoff , a PhD from the Stanford university-had an inspired thought. Why make a set of chips just for one application? Why not generalize the design so that other computing tools could be built with the same generalized design?In the process ,Dr. Hoff also suggested that the composite elements of a computer , as it was then understood- a unit to do the arithmetic (ALU-Arithmetic-Logic-Unit), a small memory and input and output interface – could be combined in a

	single slab of silicon.. When Busicom had cash flow problems, INTEL bought back the design it had created and slipped the product into its own catalogue. Intel µP 4004 was sold to Defense Establishment in India at a cost of Rs 7,000.00. Intel's Indian Agent was Hyderabad based Electronics International which was later named as Microelectronics International]
	Gordon Moore predicted that number of transistors on integrated circuits(a rough measure of computer processing power) will double every 18 months at a minimum cost. It became a self fulfilling prophecy.Moore's Law has become a yardstick of our progress as we harness the cunning of NATURE's design strategies.
1971	Electronic Mail or e-mail is invented by Ray

	Temilinson.
1972	first 8 bit µP 8008 introduced by Intel.
1977	First Personal Computer was introduced, named APPLE II , at a price of $1300.00 .It had 4kB RAM. It had a software VisiCalc the best spread sheet package of its day. This was formally the **FOURTH GENERATION COMPUTER.**
1978	Intel introduces 8086 µP Chip which had 16bit data word and 16bit address word .
1981	IBM introduced its own brand of Personal Computer PC5150 based on 8088µP Intel Chip. Clock Rate was the same as that of APPLE II but memory was much bigger at 40kB. GREEN SCREEN, 2x5” Floppy Drives for loading OS. IBM PC design accounts for 90% PCs. APPLE PC design accounts for 10% PCs.
1982	Intel introduces 80826µp

	Chip (16 bit data bus and 24 bit address bus and 68 pin package).
1986	80386(Intel) and 68030(Motorola) introduces 24 bit address bus.
1988	First Trans Atlantic Optical Fiber Cable, TAT8, used in place of coaxial cable.Coaxial Cable can carry 3,000 simultaneous phone calls. In contrast Optical Fiber cable can carry 37,500 simultaneous telephone calls with the present state of art of electronics. As the switching speed increases, the number of simultaneous calls increases exponentially.
1993	PENTIUM I introduced . it had $3.1x10_6$ CMOS(Complementary Metal Oxide Semiconductor Field Effect Transistor) placed on one chip .

- it could perform 90MFLOPS, 5 TIMES THE

SPEED OF 80486.

- data bus 32 bits and address bus 32 bits
- clock rate 1500 times that of 4004 .

2001

Pentium IV chip introduced;

- Nano-electronics chip;
- It has 42 million transistors packed in an area of 1 cm by 1 cm;
- System Clock Rate is 2GHz;
- 20,000 more powerful than 4004 µP Chip;
- Cost Rs 15,000.00

2003

APPLE introduces Power Mac G5 PC.

- 64 bit data word and 64 bit address word;
- 18 PETA Memory locations can be addressed;
- System Bus operates at 800 MHz;
- System Clock operates at 2GHz, dual processors;
- 80 to 160GB hard disc;
- 256MB to 512MB high speed RAM;

From here onward we have 64 bit computing era starts.

2004

“ Project Columbia” a $50 million super-computer built by SILICON GRAPHICS is introduced;

- it is powered by 10,240 processors in 20 units;
- it is Intel Itanium chips;
- it performs 43 Tera FLOPS as compared to the world record of 36TFLOPS

performed by IBM's BlueGene/L super computer;

- This is a ten fold increase in NASA's supercomputing capacity;
- Applications in hurricane tracking, weather modeling and earth's interior

imaging;

2005	GeForce 7800 GTX (latest Graphics Processing Unit or GPU) introduced by US-based nVIDIA Corporation for Video Games;

- 302 million CMOS on one 1cm x1cm chip;
- Performs 202 GFLOPS;
- 10% less power consumption ;
- 50% quieter
- twice as powerful as its processors 6800 but it costs the same as before that is

Rs. 32,000.00 ;

- Market for video games is one of the fastest growing sectors in Computer

	Industry;
2005	Dual cores Microprocessor chip are introduced;Intel introduces Pentium “Extreme Edition” 840 with Clock Rate = 3.2GHz and with billon transistors . This is dual core as well as it has a software called‘hyperthreading’(HT). This is a software which enables one to work with doublevirtual processors. Thus a dual core with HT will allow application developers to write software that can carve up the task between four processors; • Intel introduces Xeon for servers and Pentium M for laptops; • AMD introduces Opteron 800 series for servers and workstations; • AMD introduces Athlon 64 X2 for consumers and business PC’s;
2008 October	GeForce 9400M is a single chip solution for graphical processing platform for portable applications. It has

	computing rate of 55GFlops (Floating Point Operations), the most powerful integrated graphics processor in the market in 2008. It has 6 cores operating in parallel. But much less power hungry. New MacBook family will be fuelled by these chips.
2008November	Intel and AMD have launched a native quad processor unlike the earlier version where 2 dual core chips had been combined in the same package. Intel has launched ‘Nehalem’,the i7 processor using 45nm technology. AMD has launched 45nm Opteron. Both achieve the 4 cores on the same silicon slab. This has double the Memory Bandwidth. These conserve energy by cutting down power when the device is idle.
2009	Interplanetary Internet Protocols are tested by

	using computers on the International Space Station. Eventually there will be an internet connectivity among Earth, multiple probes, rovers, orbiters and space crafts exploring the solar system. This will be the basis of Interplanetary Internet. The network's key technologies are called Delay Tolerant Networking(DTN) protocols. The problem in deep space internet is the huge delay and occultation of the nodes. To overcome this problem "store and forward" technique will be used. The node will save the data and then passes on when the link is again established. A third test is yet to be carried. In this a security protocol and a new file-transfer protocol will be used.
2010	Pacific Unity will lay multiple fibers. Total

capacity; 7.68 terabits per second.

ROAD MAP OF SUPERCOMPUTER.[‘The Data’, IEEE SPECTRUM, APRIL 2009]

Table I.3. Road Map of Super Computer.

Year	Name	No.of processors	Capability	Cost
1976	Illiac IV	64	TeraFLOPS	480kFLOPS/ million dollar
1976	Cray 1	1	TeraFLOPS	17.78MFLOPS/ million dollars
1988	CRAY Y-MP	8-vectorprocessor	TeraFLOPS	111.5MFLOPS/ million dollars
1997	ASCI RED	4510	TeraFLOPS	18.18182GFLOP million dollars
2002	EARTHSIMULATOR	5120	TeraFLOPS	17.5GFLOPS/ million dollars
2004	BLUEGENE/	65536	10TeraFLOPS	2.8TFLOPS/

	L			million dollars
2007	PLAYSTATION3 CLUSTER†	8playstation		375TFLOPS/ million dollars
2008	NVIDIA TESLA*	960cores		439.1 TFLOPS/ million dollars
2008	ROADRUNNER‡	19440	10sTeraFLOPS	8.3TFLOPS/ million dollars

*Based on the cost of building your own Tesla supercomputer out of four Tesla C1060 units. Complete instructions are available at nvidia.com

† Seven processors per node at 150GFLPS per Unit for a single processor(11 processors per node for a dual processor). $400-plus per node.

‡ 6480 Opteron CPUs and 12960 Cell processors.

Supercomputers are massively parallel chips can be used to calculate protein folding, predict climate change and crack the encryption of hitherto-secure Web sites.

Nvidia advertises its new workstation, the Tesla, as a “personal supercomputer”. It clusters 4 Nvidia C1060 processing boards, each of which unites 240 graphic cores to process instructions at nearly teraflop speed.

Tesla does single-precision floating-point calculations using 8-bit bytes.

Roadrunner uses 64-bit floating integers.

Table I.4.Evolution of Microprocessor Chip

Name	Technology	Data & Address Size	Packing Density	Number of Pins	Clock Rate	Instructions Per Second	Year & Price.
Intel 4004	10µm	4-bit	2300PMOS	16	740kHz	92kIPS	1971 Rs 7,500.00
Intel 8080	3 µm	8-bit	6000NMOS	40	2MHz	0.64MIPS	1977
Intel 8086	3 µm	16-bit	29,000 HMOS	40	5 MHz	0.33MIPS	1978
Intel 80286	1.5 µm	Data 16b Address 24b	134,000 HMOS	68	6-12.5 MHz	1.8 MIPS	1982
Intel 80386	1 µm	32-bit	275,000 CMOS	132	16 MHZ	5 MIPS	1985

Intel 80486	800nm	32-bit pipelined	1.2 million CMOS	168	25 MHZ	20 MIPS	1989
Pentium	600nm	Superscalar architecture Data 64b Address 32b Memory Size 2^{32} = 4GB	3.1 million CMOS	?	60 MHZ	100 MIPS	1993
Pentium II	350nm	Same	7.5 million CMOS	370	233 MHZ	400 MIPS	1997
Pentium III	250nm	same	9.5 CMOS	370	450 MHZ	1000 MIPS	1999
Pentium IV	180nm	Data 64b Address 64b	42 million CMOS(Nano electronics)	478	2.2 GHz	5GIPS	2001

MOORE’s LAW: The packing density of CMOS on μP Chip would double every two years leading to an exponential growth in the complexity and speed of μP Chip.

Table I.5. Evolution of Computers.

Generation	Components	Architecture	Logic	Programming	Language
	Von-	Boolean	Algorithm	Fortran,Basic,Algol	

		neumann, sequential				
5th	Robotics, Data Based Medical Diagnostic Tools	Artificial Intelligence Machines	Data-flow, Systolic, Pipeline	Predicate	Heuristic	Prolog, Lisp

Table I.6. Evolution of Local Area Network

Year	Standard	Name	Data transfer rate
1973	802.1	OSI (Open System Interconnection)	10Mbps
	802.2	Logical Link Control	10Mbps
1983	802.3	CSMA/CD (Carrier Sense Multiple Access/ Collision Detect)	10Mbps

	802.4	Token Bus	100Mbps
	802.5	Token Ring	100Mbps
1999	802.11	WLL (Wire Local Loop)	1 Gbps
2001	802.13	Wi-Fi (Wireless internet Access	

Table I.7. Development in submarine cable. (Sources: Telegraphy Research: *"History of the Atlantic Cable and Undersea Communications"* **by Bill Glover,Wikipedia).**

Region	Submarine Cable %BW utilization	LIT Submarine Cable Capacity
Asia	69%	
Middle East	73%	
Oceania	56%	
Europe	52%	
US & Europe	51%	
Latin America & Caribbean	77%	
Africa	69%	
Transatlantic Cable		15%

US Latin America	37%
Transpacific	21%
Europe Asia	91%
Intra-Asia	31%

Submarine Cable Boom: In 2001 US $13.5 billion was spent. Due to Dot.com bust, only $2billion was spent during four lean years from 2004 to 2007. Even the during the lean periods, global demand never slaked growing at an average compound annual rate of 54% from 2002 to 2008. In Latin America and the Caribbean, traffic grew the most at more than 75%.

METALFE LAW: The value of a NETWORK grows as the square of the numbe of users.

50 years Journey of IC Technology_Appendix II
This gives the steps in photo-lithography technique.
How the window pattern is made on the Photo-resist and how the window is opened in SiO2 layer.

Appendix II.

Steps in Optical Lithography

(Film-and-chemistry photography)

Blank Patch on a Silicon Wafer(12¨- Ø)

↓

Silicon Wafer is held on the chuk of a spinner by vacuum suction

↓

Next it is coated with photoresist and spun at a high r.p.m. to ensure uniform layer of photo resist across the wafer

↓

The correct mask is placed in contact with the photo resist layer

↓

The photo resist is exposed to Ultra-violet light for 1 minute through the mask

↓

The exposed region gets polymerized and hence hardened whereas the unexposed portion is soft and can be dissolved away

↓

Next wafer with the exposed Photo-Resist coating is dipped In a developer

↓

The developer dissolves away the soft portion leaving a window pattern on the hardened portion of the photo-resist layer

↓

Next the wafer along with the the window pattern printed on the photo resist is dipped in HF acid which is an etchant of SiO2. HF acid is a very strong solvent. It is placed in Teflon beaker and Teflon twigger is used for handling the wafer. HF acid must never come in touch with our skin. If it comes it will reach the bone and dissolve it. So the personnels will have to use Teflon gloves and suitable aprons for safety as well as for cleanliness

↓

The window pattern of the hardened resist is stripped off and the wafer is ready for diffusion or ion implantation or for metallization

↓

Hundreds of identical microchip pattern is inscribed on the wafer by optical photo-lithography, plasma etching and a combination of ion-implantation or diffusion

↓

Sliced into microchips using a diamond tipped scriber

↓

Individual microchips also referred to as dies are mounted on headers

↓

Gold wires are connected from the bonding pads on the die to the external leads on the header by thermo-compression bonding or by Ultra-sonic bonding

↓

Hermetically sealed

↓

Sent for applications

(laptops, thumb drives, cell-phones and GPS)

50 years Journey of IC Technology_Appendix III_Alternatives of NGL
Appendix III describes the next generation technologies being developed as Optical Lithography reaches its resolvable limit of minimum feature size of 30nm.

Appendix III

Five Alternatives of Next Generation Lithography

[“Next Generation Lithography”by Eric J. Lerner, June 1999,American Institute of Physics,pp 18 to 21.]

In four decades following the invention of IC chip, speed has doubled every 3 years and smallest feature size has scaled down by 2 every 6 years in accordance with Moore’s Law. This accounts for half the increase in Instruction per second executed by a given Computer of a given technology. The other half of the increase in IPS(instruction per second) has come about by the improvement in the architecture of microprocessor chip. Von Neumann sequential architecture has evolved into parallel architecture such as pipelining, systolic and data flow architecture. This has enabled to carry out complex instruction in reduced number of machine cycles. With the downward scaling of the devices, the clock speed has increased and simultaneously the number of machine cycles utilized to carry out an instruction has decreased. Both these factors increase the computational speed of a given computer of a given technology.

As devices have been scaled down, the line width has decreased. This has required that the wavelength of the light used in the photo-lithography is correspondingly decreased according to the formula:

W = minimum feature size = (according to Rayleigh Equation) = $k.\lambda/(NA)$

where NA(numerical aperture) = $n.\mathrm{Sin}(\alpha) = d/(2f)$;

n = index of refraction of the medium surrounding the lens and α = acceptance angle of the lens system;

d = aperture diameter and f = focal length.

k = resolution factor.

Lithography underlies the entire I.C. technology. I.C chips are created by alternatively etching the oxide layer and creating a window pattern, through the window pattern carrying out the diffusion of dopents or implanting the dopents. Etching out a window pattern involves a pattern of photo resist imprinted on the chip through Photo Lithography Technique. The photo-lithography involves masks. Masks consist of chromium lines laid on quartz substrate. They are generated by computer controlled electron beams which lay down the pattern of chromium lines. The masks are 4 times larger and wider than the circuit pattern actually required. By optical projection the mask pattern is reduced 4 times to the correct size. This enables the mask generation more accurate and technically viable. In the mask the minimum feature size is 4 times the actual size.

Below 30nm we have to go for Next Generation Lithography. The five alternatives are:

1. Extreme Ultra-Violet Lithography;
2. X-Ray Proximity Lithography;
3. Projection Electron Beam Technology;
4. Ion Projection Lithography;
5. Improved Optical Lithography.

1. Extreme Ultra-Violet Lithography(EUV lithography):

The equipment set-up has been described in Part 3 of the main text in Figure 3 . The description of the setup is as follows:

Laser produced plasma

↓ radiation

Irradiates the supersonic jet of Xenon Gas

↓reradiates

EUV(13.4nm equivalent to 100eV energy)

↓

Complex Optical System

↓

Through the optical system, EUV is reflected to Reflective Mask with pattern

↓

From the Reflective Mask, EUV is reflected to the substrate's photo resist coating with 4 times reduction creating the correct size chip pattern in the photo resist for further processing

1. X-Ray Proximity Lithography:

Here the wavelength used is Soft X-Ray of wave length one order of magnitude lower i.e. of 1.34nm $= 13.4A^{\circ}$ (1000eV)

Synchrotron Radiator is used as X-Ray Source

↓

Thin stripes of radiation generated

↓

The stripe of radiation is bounced off a low-incidence angle oscillating mirror

↓

Bounced off beam is scanned across the mask area

↓

Chip pattern is imprinted on the Photo Resist coating of the wafer

Here the mask pattern is of the same size as that of the pattern imprinted on the wafer. Mask is made of SiC , 2μm thick and 5 cm wide. The mask is kept thin to make it transparent to Soft X-Ray. These are fragile and because the mask pattern is of the same size as the actual pattern hence electron beam generation of the pattern on the mask becomes difficult. Soft X-Ray has deeper penetrating power hence existing photo-resist suffice X-Ray proximity lithography requirement.

1. Projection Electron Beam Lithography:

Electron's wave-particle duality allows its wave nature to be utilized for electron beam lithography where electron wave corresponds to a wave length λ. This wavelength is determined by its linear momentum $p = mv$.

From deBroglie wave particle duality: $\lambda = h/p = h/(mv)$;

If an electron beam is accelerated through 100kV it acquires a kinetic energy of 100keV.

Therefore $(1/2)mv^2 = 100\text{kev} = (100 \times 10^3 \times q)$

Therefore $mv = \surd(200 \times 10^3 \times q \times m)$ where m = mass of an electron.

Therefore $\lambda = h/[\surd(200 \times 10^3 \times q \times m)]$

Substituting the numerical values of h, q and m:

$\lambda = 3.9$ pico meter.

This wave length is more than sufficient for resolving features of nm size. By using differentially scattered electron beam, electrons through the transparent area exposes the photo resist whereas in pattern area the electron beam is prevented from transferring its energy to the photo resist. This system is called “Scattering with angular limitation projection electron beam lithography” known by its acronym “SCALPEL”. Diagram Figure III.1 describes the SCALPEL system.

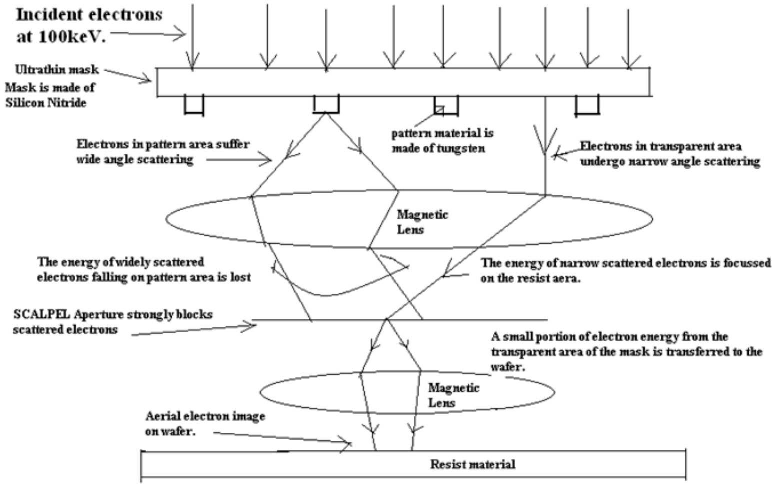

Incident electrons at 100keV.
Ultrathin mask
Mask is made of Silicon Nitride
Electrons in pattern area suffer wide angle scattering
pattern material is made of tungsten
Electrons in transparent area undergo narrow angle scattering
Magnetic Lens
The energy of widely scattered electrons falling on pattern area is lost
The energy of narrow scattered electrons is focussed on the resist aera.
SCALPEL Aperture strongly blocks scattered electrons
A small portion of electron energy from the transparent area of the mask is transferred to the wafer.
Magnetic Lens
Aerial electron image on wafer.
Resist material

Figure III.1 The schematics of SCALPEL.

Here 4-to-1 reduction is allowed hence the mask is conveniently generated but fragility of the mask remains a problem.

1. Ion Projection Lithography:

This is the least developed of all the NGL. It uses accelerated hydrogen or helium ions instead of electrons and focuses them using electrostatic field. This has a mask which has holes punched for allowing the ions to pass through. The holes are the transparent equivalent of the conventional optical masks. Hence this is stencil like mask. The stencil mask does not allow closed path of conductors . This is the donut problem.

1. Retaining Optical Lithography:

Next step in Moore’s scaling aims at 100nm and 70 nm generation chips. Here Excimer Laser source also known as Flouride Lasers will generate 157nm monochromatic light. Calcium Fluoride lenses could be used for focusing. Resolution enhancement could make optical lithography viable for 100nm and 70nm generation chips.

50 years Journey of IC Technology_Appendix IV_Band Gap Narrowing and its effect on Emitter Injection Efficiency

Appendix IV describes the effect of band gap narrowing in degenerate semiconductor.

Appendix IV

Band Gap Narrowing(BGN) and its effect on Emitter Injection Efficiency.

In homojunction:

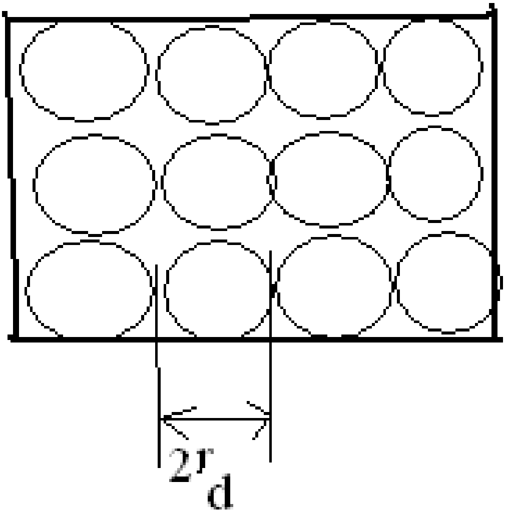
$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{n_{BO}}{p_{EO}}\right)$$

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{n_{iB}^2}{N_{A|B}}\right)(\frac{N_{D|E}}{n_{iE}^2})$$

To obtain high gain, Emitter needs to be heavily doped and we need thin Base Region.

$$\frac{n_i^2}{1} = \frac{N_C N_V Exp(\frac{-E_g}{kT})}{1}$$

Due to heavy doping in Emitter, degeneracy is introduced which leads to Band Gap Narrowing(BGN) by ΔE_g . The radius of the fifth electron or the donor electron orbiting donor atom is 13×10^{-8}cm = 13Angstrom = 13Aº.



Orbitals of the donor atom electrons for the critical concentration N_{crit}

Figure 1

Figure IV.1.Calculation of critical packing density where donor electrons wave function of donor atoms start to overlap.

$N_{crit} = 1/V_d = 1/(2r_d)^3 = 1.7 \times 10^{19}$/cc.

When the doping level approaches this critical level, energy level corresponding to donor atoms no more remain DISCRETE but they become a continuum in effect extending the conduction band and narrowing the Forbidden Energy Band Gap which is known as Band Gap Narrowing(BGN). This is a degenerate semiconductor and the Fermi-level lies at the edge of the conduction band or within the conduction band as it is in metal. Hence a degenerate semiconductor is semi-metal. The energy band diagram of degenerate and non-degenerate semiconductor is given in Fig.IV.2.

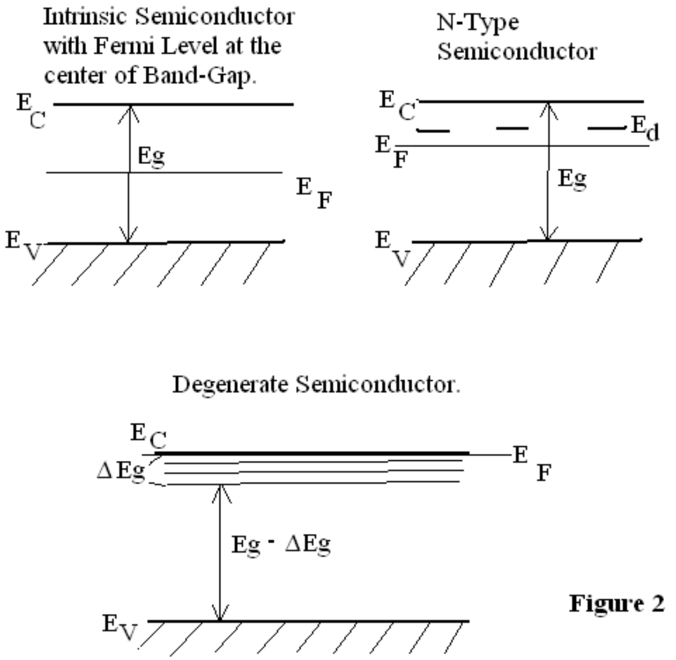

Intrinsic Semiconductor with Fermi Level at the center of Band-Gap.
E_C
Eg
E_F
E_V
N-Type Semiconductor
E_C
E_d
E_F
Eg
E_V
Degenerate Semiconductor.
E_C
ΔEg
E_F
Eg - ΔEg
E_V

Figure 2

Figure IV.2. The Energy Band Diagram of Pure, doped and heavily doped semiconductor.

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{n_{iB}^2}{N_{A|B}}\right)\left(\frac{N_{D|E}}{n_{iE}^2}\right)$$

IV.1

$$\frac{n_{iE}^2}{1} = \frac{N_C N_V Exp(\frac{-(E_{go} - \Delta E_g)}{kT})}{1}$$

$$\frac{n_{iB}^2}{1} = \frac{N_C N_V Exp(\frac{-(E_{go})}{kT})}{1}$$

$$\frac{(n_{iE}^2)}{1} = \frac{n_{iB}^2 Exp(\frac{\Delta E_g}{kT})}{1}$$

IV.2

With BGN the intrinsic carrier concentration increases

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{n_{iB}^2}{N_{A|B}}\right)(\frac{N_{D|E}}{n_{iE}^2})$$

Substituting Eq.IV.2 in Eq.IV.1 we get

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{1}{N_{A|B}}\right)(\frac{N_{D|E}}{Exp(\frac{\Delta E_g}{kT})})$$

IV.3

Therefore

$$N_{Deff} = N_D Exp(-\frac{\Delta E_g}{kT})$$

IV.4

The empirical relationship for BGN is:

$$\Delta E_g = E_{ref}\left\{ln\left(\frac{N_D}{N_{ref}}\right) + \left[\ln\left(\frac{N_D}{N_{ref}}\right)^2 + 0.5\right]^{0.5}\right\}$$

Where E_{ref} = 0.009eV and N_{ref} = 10^{17}/cc for typical cases.

Table IV.1. ΔE_g vs N_D .

Emitter Doping Density (number per cc) N_D	BGN ΔE_g (eV)
10^{20}	0.12466
10^{19}	0.083
10^{18}	0.0424
10^{17}	6.36×10^{-3}
10^{16}	1meV

Table IV.2. N_{Deff} vs N_D.

N_D (number per cc)	ΔE_g (eV)	N_{Deff} (number per cc)
10^{20}	0.12466	8.258×10^{17}
10^{19}	0.083	4×10^{17}
10^{18}	0.0424	1.96×10^{17}
10^{17}	6.36×10^{-3}	8×10^{16}
10^{16}	1meV	1×10^{16}

For a doping level of $N_D = 10^{20}$/cc, $\Delta E_g = 0.12$eV.

This gives an effective doping level of $N_{Deff} = 10^{18}$/cc.

The classical formula for current gain without BGN is:

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{N_{D|E}}{N_{A|B}}\right)$$

If we take the following data: $(N_D)_E = 10^{20}$/cc, $(N_A)_B = 10^{17}$/cc, $W_E = W_B = 1\mu m$ and $D_{pE} = 1.25$ $(cm)^2$/sec and $D_{nB} = 20$ $(cm)^2$/sec then $\beta_F = 1.6\times10^4$;

But taking into account of BGN, $\beta_F = 160$.

50 years Journey of IC Technology_Appendix V_Polysilicon Emitter BJT
This Appendix is based on the patent filed in US for a simplified method of fabricating poly-silicon emitter BJT.

Appendix V

Analysis of Polysilicon emitter BJT .

In Polysilicon emitter BJT (PET), a polycrystalline Si Film lies between Metallic Emitter Contact and Silicon Proper. By ion-implantation, the Polycrystalline Silicon Film is doped with Donor Arsenic Ions. By subsequent annealing, shallow diffusion of Donor Arsenic atoms from N+ Poly-Silicon Source takes place leading to a shallow layer of N+ in Si proper contacted by a similarly doped N + Poly-Silicon thin film. This structure is known as Polysilicon emitter BJT.

This is suitable for shallow EB Junction;

Compatible with self-aligned processing. This minimizes parasitic Resistance and Capacitances leading to faster response time.

Short Circuit Current Gain (β_F) in PET is considerably improved as compared to that of shallow all crystalline emitter of equivalent thickness.

The physics of improvement of Short Circuit Current Gain (β_F) in PET.

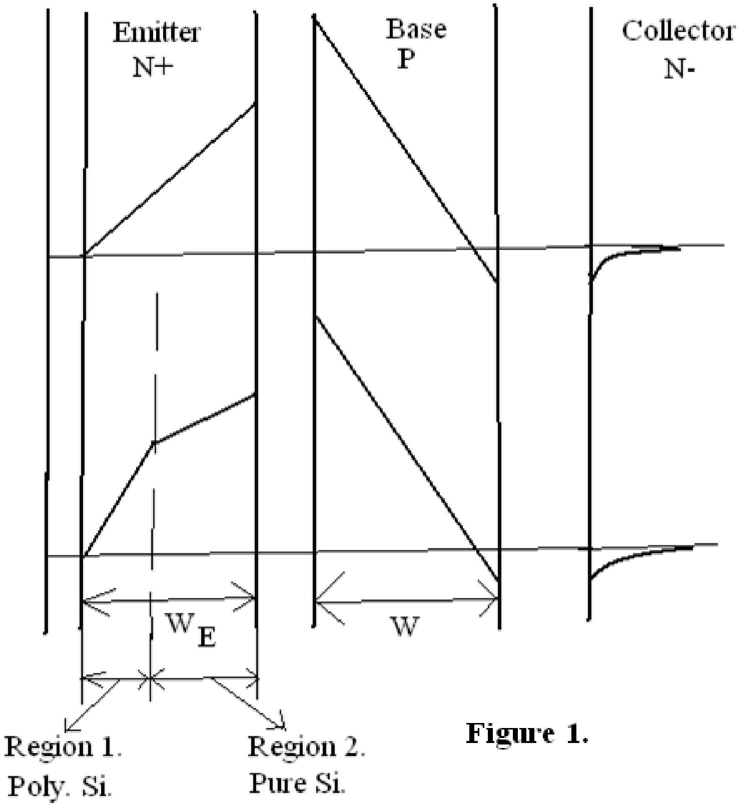


Figure V.1. Minority Carrier Concentration in a shallow EB Junction case 1. All pure single crystal structure and case 2 in Poly Si Emitter BJT.

Since $W_E << L_{pE}$, therefore there is negligible recombination through out the Emitter. At the interface of Poly Si(region 1) and Pure Si(Region 2), continuity of carrier concentration and continuity of diffusion current requires that:

$$qDe1^{\frac{d(pE1)}{dx}}$$

=

$$qDe2^{\frac{d(pE2)}{dx}}$$

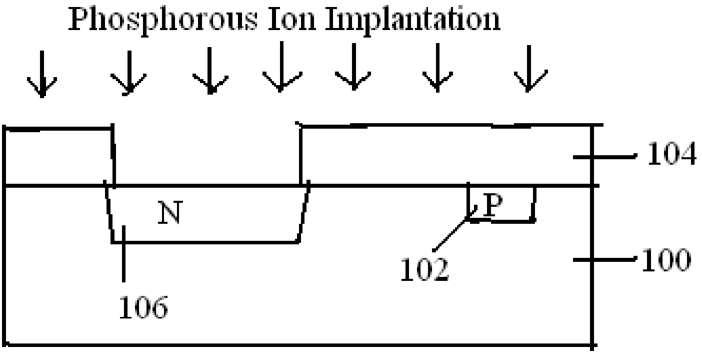
$$\frac{d(pE1)}{dx} = \frac{De2}{De1} \times \frac{dpE2}{dx} = \frac{\mu e2}{\mu e1} \times \frac{dpE2}{dx}$$

Therefore carrier concentration gradient in poly Si region is several times the carrier concentration gradient in Pure Silicon Emitter because mobility of hole in Pure Single Crystal Silicon region is several times greater than that in Poly Si on account of domain scattering in Poly region. Hence hole current crossing EB junction into Emitter is considerably reduced leading to a near unity Injection Efficiency. This leads to a much better α and hence a much better β . Now doping of base can be increased leading to a reduced base spreading resistance and hence better Figure of Merit and better Frequency Response without losing on current gain.

Process Steps of Poly Si Emitter BJT US Patent5736447. (P+NP)

www.patentsform.us/patents/5736447/fulltext.html

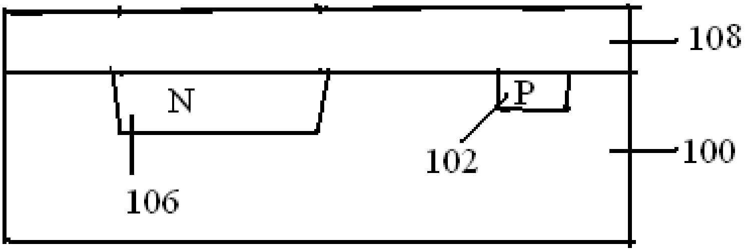
Step 1. Ion-Implantation of N type Base in P type substrate. Base Region is in space-apart relationship to the collector region.



100 - P type substrate, 102- Collector P Type
104- Kodak Photo Resist(KPR) film over the substrate with a window for ion-implantation of N-type Base.
106- Ion Implanted Base Region(N-Type).
Phosphorous is ion-implanted to obtain the BASE.
Implantation Voltage 136 to 180 KeV.
Dose 3*10^13 to 1*10^14 ions/cm^2

Figure 2.

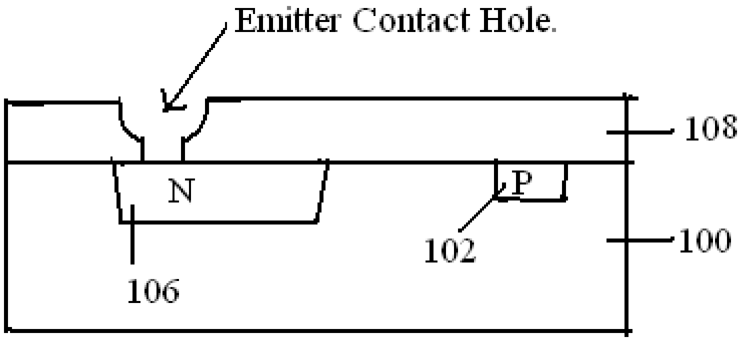
Step 2. Stripping of exposed KPR layer(104) and deposition of FIRST Insulation Layer.



Polymerized KPR is stripped by stripper and the surface of the P Type Substrate is covered with the First Insulating Layer.
First Insulating Layer could be Low Temperature Oxide (LTO) or High Temperature Oxide (HTO) or double layer LTO+Borophosphosilicate glass(BPSG) or HTO+BPSG.
108- First Insulating Layer deposited to a thickness of 2000 to 6000 Angstrom.

Figure 3

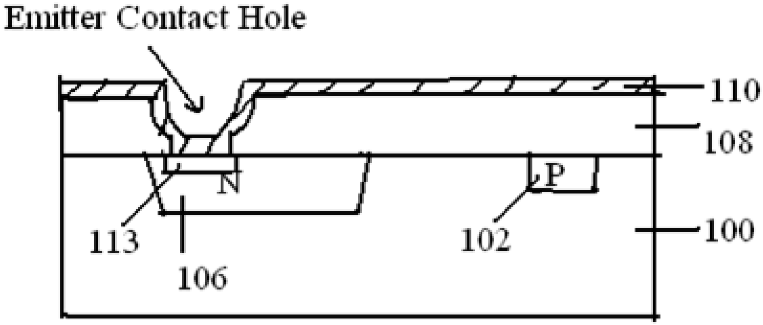
Step 3. Etching out of Emitter Contact Hole. The oxide film (108) is first isotropically etched and then anisotropically etched or we could do anisotropic etching directly



108- First Insulating Layer deposited to a thickness of 2000 to 6000 Angstrom.

Figure 4

Step 4.Deposition of first conducting layer of thickness 500 to 2000 Angstrom. The first conducting layer is Poly-Silicon layer or in-situ polySi. P-type impurities namely Boron Ions are Ion-Implanted into the first conducting layer. Ion-Implantation energy is 50 to 100keV and the dose is 2×10^{15} - 6×10^{17}ions/cm^2. Ion-implanted poly-Si layer acts as the source for shallow diffusion of P-type impurities in N-type Base to form a shallow self aligned Emitter in Pure Single Crystal Silicon Substrate. This shallow diffusion is achieved by rapid thermal annealing(RTA) at a temperature above 600ºC in a nitrogen atmosphere at a temperature of 800ºC - 950ºC.



108- First Insulating Layer deposited to a thickness of 2000 to 6000 Angstrom.

110- First conducting layer deposited over (108) first insulating layer and over (106) exposed Base region.

113- Shallow Emitter achieved by Rapid Thermal Annealing of (110)P+polySi first conducting layer.

Figure 5

Step 5. Base and Collector contact holes etched out in first conducting layer(110) and in first insulating layer(108) to Base(106) and to Collector(102). This is achieved by isotropic and then anisotropic etching or only by anisotropic etching.

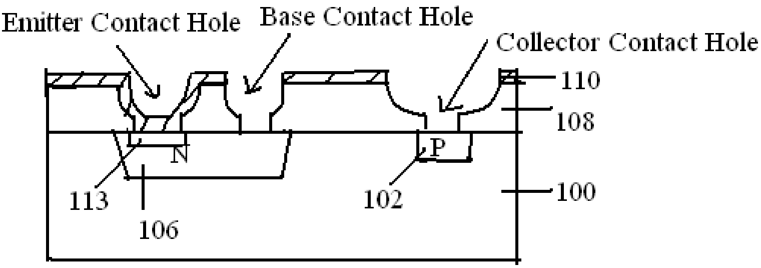
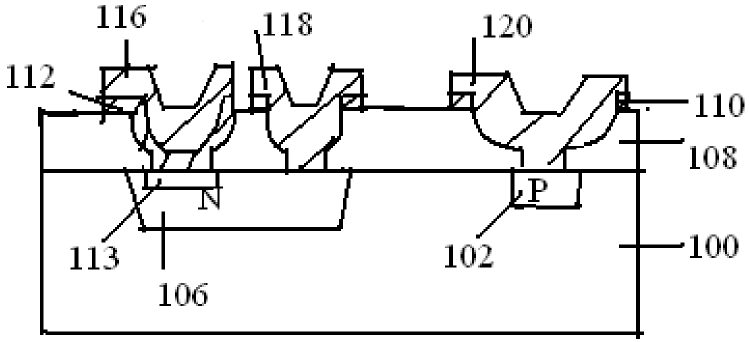


Figure 6

Step 6. Second Conducting Layer is deposited over 110 and to fill the contact holes at Emitter, Base and Collector. It may be Poly-Si or poly-Si in-situ or Metal Silicide. Second Conducting layer plus underlying conducting layer are patterned into Emitter Contact Pad, Base Contact Pad and Collector Contact Pad.



112- Poly-Si Emitter;
116- Emitter Contact Pad in contact with polysilicon emitter 112;
118- Base Contact Pad in contact with 106 Base Region;
120- Collector Contact Pad in contact with 102 Collector Rgion;

Figure 7

Directly quoted from the Patent:

“The above-described method of the present invention results in a BJT having a thin or shallow emitter region 113 which effectively minimizes the base-emitter junction capacitance, without the need for additional photolithographic steps required by the presently available technology. In this regard, the polysilicon emitter 112 is formed simultaneously with the patterning of the first and second conductive layers and is self –aligned with the resultant emitter contact. Moreover the necessity of the oxide layer 38 of the conventional method for manufacturing a BJT is eliminated. Hence this method is less complex and less costly than the conventional method”.

50 years Journey of IC Technology_Appendix VI +VII_Atomic Layer Deposition +FermiLevelPinnning

Appendix VI and Appendix VII deal with two major problems of CMOS Technology namely how to achieve smooth and even oxide layer for improved yield and how to prevent Fermi Level Pinning.

Appendix VI.

Atomic layer Deposition

Dielectric Layers were deposited by Reactive Sputtering or by Metal Organic CVD. This left unevenness on the surface which caused charge trapping. There were charges stored at the interface too. These charges altered the behavior of MOS capacitance from discharge cycle to discharge cycle. So a deposition method had to be adopted which allowed dielectric deposition in controlled manner atom layer by layer.

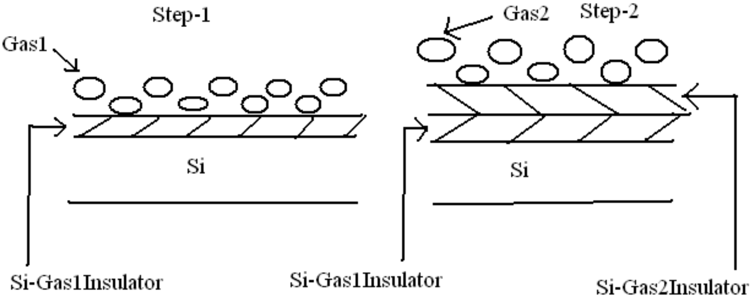


Figure A

Figure VI.1. Methodology of Atomic Layer Deposition.

In Figure VI.1 the steps taken for Atomic Layer Deposition is illustrated. Gas1 reacts with bare Silicon surface to form a single atomic layer of insulator Si-Gas1. As soon as one layer is completed Gas1 stops reacting with Si because no bare Si is in contact with Gas1.

In step 2, Gas2 is chosen which is reactive with dielectric1 or Si-Gas1 insulator. Gas2 reacts with dielectric layer1 to form dielectric layer2 (Dielectric1-Gas2 insulator). As soon as dielectric layer2 covers the whole of dielectric1, reaction stops. Thus each reaction is terminated at the end of layer deposition. In this way we achieve layered gate insulator which is controllable down to the width of a single atom.

This produces much smoother dielectric hence the charge trapping is prevented . MOS produced is reproducible and stable in operation.

Appendix VII.

Fermi-level pinning and its prevention.

(academic.brooklyn.cuny.edu/physics/tung/Schottky/index.htm)

At Metal-ntype-Semiconductor interface we realize Schottky Barrier Diode. This exhibits a rectifying contact characteristic hence it has an in-built Schottky Barrier Height(SBH) as illustrated in Figure VII.1.

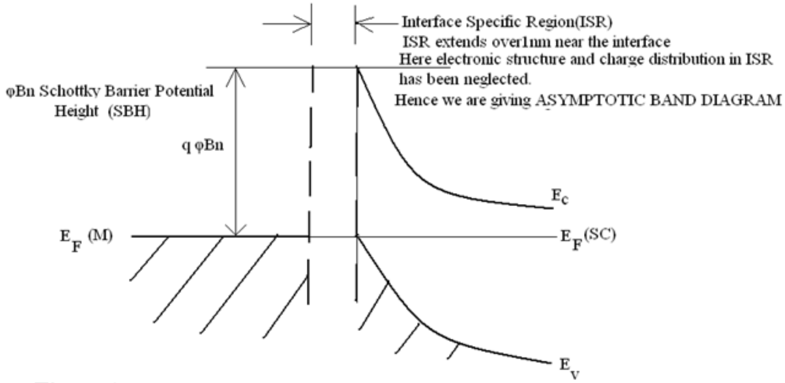

Interface Specific Region(ISR)
ISR extends over1nm near the interface
Here electronic structure and charge distribution in ISR
has been neglected.
Hence we are giving ASYMPTOTIC BAND DIAGRAM
φBn Schottky Barrier Potential
Height (SBH)
q φBn
E_c
E_F (M)
E_F(SC)
E_V

Figure 1

Figure VII.1. Asymptotic Energy Band Diagram of Metal-Semiconductor Interface and illustration of Schottky Barrier Potential Height.

It was found experimentally that SBH is insensitive to the Metal Work Function. This was defined as Fermi-Level pinning (FL pinning). It generally occurred at the mid-band gap of Semiconductor. This was always the case with polycrystalline MS interfaces and it was quite counterintuitive. In 1980s a few high quality, single crystal MS interfaces prepared and SBH measured. SBH was found to be sensitive to orientation/structure of of MS interface. By spatially-resolved SBH measurement technique notably by Ballistic Electron

Emission Microscopy(BEEM) it was found that SBH are inhomogeneous at polycrystalline MS interfaces and structure dependent at single crystal MS interfaces.

At the turn of the century when dipole associated with chemical bonding at MS interfaces was modeled using established methods borrowed from Molecular Physics it was shown that FL pinning was a natural consequence of interfacial bonding.

If minimization of total energy is applied

1. then it results in structure dependence of SBH at epitaxial interfaces and
2. SBH inhomogeneity at polycrystalline interfaces.

By this theory FL pinning at polycrystalline MS interfaces and the pinning position at mid-band gap

comes automatically.

This model is being further refined.

50 years Journey of IC Technology_Appendix VIII_Mobility Enhancement in strained Silicon

This Appendix describes the method of creating tensile and compressive stress in Silicon Bulk to enhance the mobility of electron while drifting across the channel.

Appendix VIII.

Mobility enhancement in strained Silicon.

[Mobility Enhancement, the next vector to extend Moore's Law, Nidhi Mohata & Scott E. Thompson, IEEE Circuits & Devices Magazine, September/ October 2005, pp.18-23]

Geometric Scaling has been driving IC Industry till date. Since 90-nm Technology generation was introduced, off-state leakage current and power density have made scaling a difficult and challenging job. New scaling vectors were adopted to meet this challenge. At 90-nm generation, mobility enhancement through uniaxial process-induced strained Si has emerged as the next scaling vector.

The theoretical formulations of carriers in 2-D inversion layer just below the insulator Gate:

$$\mu(carrier\ mobility) = \frac{q\tau}{m^*}$$

where τ is the mean free time between two consecutive scatterings;

$1/\tau =$ scattering rate;

m^* is the conductivity effective mass.

Under strain, both m^* and scattering rate reduce leading to enhanced mobility.

Uniaxial stress always provide higher current enhancement(1.46mA/µm and 0.88mA/µm for n-channel and p-channel respectively) as compared to that produced by biaxial stress(0.85mA/µm and 0.45mA/µm for n-channel and p-channel respectively). Hence in state of art technologies, Industries have adopted uniaxial stress. There are three state-of-art techniques:

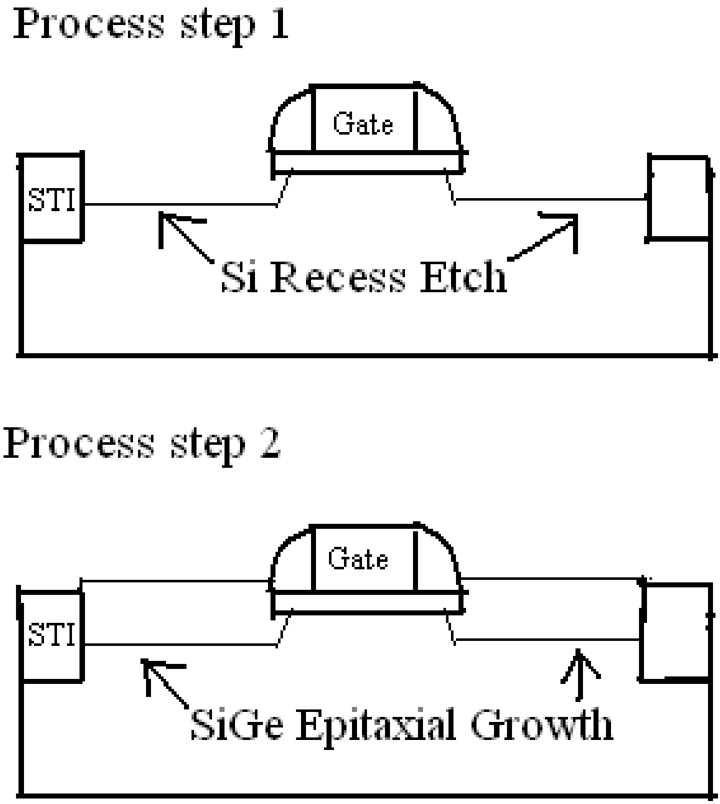
1. A local epitaxial film is grown in the source and drain regions;
2. In the second technique we use a capping layer;
3. A dual capping approach.

VIII.1. First Approach of local epitaxial film.

The process flow is shown in Figure VIII.1. As shown first Si source and drain are etched, creating an Si recess.

In next step, SiGe (for p-channel) or SiC (for n-channel) is grown in the source and drain region.

This creates uniaxial compressive stress in p-channel or uniaxial tensile stress in n-channel MOSFET. 17%Ge, gives 500-900MPa of channel stress. 60 to 90% drive current enhancements on short cannel devices(~35nm) have been demonstrated. This is greater than that achieved in multigates or high-k dielectrics.



Process step 1

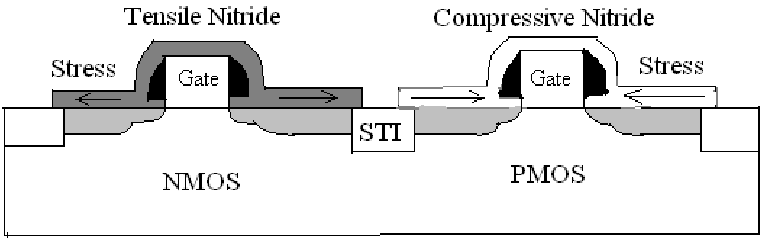
Process step 2

Figure 1

Figure VIII.1. Process Flow of local epitaxial growth.

VIII.2.Second Approach of the use of a tensile and/ or compressive capping layer.

The process flow of the second approach is shown in Figure V.2.



Dual stress liner process architecture with tensile and compressive Si Nitride capping layers over NMOS and PMOS respectively.

Figure 2

Figure VIII.2. Process Flow of Dual Capping with permanent layer.

In the second approach, capping films are introduced as a permanent layer post salicide. STI is shallow trench isolation.

VIII.3. Third Approach of capping with sacrificial layer.

In the third approach , an uniform layer of high tensile Si_3N_4 liner is deposited over the entire surface post salicidation. This is patterned and etched so that capping is removed from PMOS and is retained over NMOS. Next a compressive SiN is deposited over the entire surface and selectively patterned and etched off from NMOS. Thus Compressive Nitride is retained over PMOS. This produces comparable stresses as compared to the first approach at the same time it has reduced process complexity and integration issues.

VIII.4. Fourth Approach of Stress Memorization of Poly-Si Gate.

In the fourth approach, the following is the process steps:

1. Poly-Si gate amorphization;
2. Deposition of a high-stress SiN layer on top of the poly-Si gate;
3. Recrystallization of the poly-Si gate during source/drain anneal;
4. Removal of SiN layer.

After removal of the capping layer the memory of stress is retained in the gate as well as channel.

This is able to enhance the mobility by 10% in NMOS..

Future Scalability of Strain.

In long channel devices:

$$I_{d(on)} = \frac{\mu W C_{ox}}{L} (V_g - V_t - \frac{V_d}{2}$$

)

$$V_d$$

$$I_{d(sat)} = \frac{\mu W C_{ox}}{2L} \left(V_g - V_t\right)^2$$

As we scale down, short channel effect start showing up until MOSFET carrier transport becomes ballistic. Under Ballistic transport saturation drain current is described by the following equation:

$$I_{d(sat)} = W C_{ox} < v(0) > \left(V_g - V_t\right)$$

Where $<v(0)>$ is the average velocity at the source;

$$C_{ox}\left(V_g - V_t\right) = is\ the\ charge\ density;$$

In the limit, where channel length becomes zero and electrons or holes are ballistically thrown across the pinched off region, there $<v(0)>$ = unidirectional thermal velocity = v_T =

$$\sqrt{\frac{2 k_B T_L}{\pi m_t^*}}$$

$$m_t^* = transverse\ effective\ mass\ of\ electron\ or\ hole\ as\ the\ case\ may\ be;$$

$$T_L = lattice\ temperature;$$

Therefore

$$I_{d(sat)} = WC_{ox}\sqrt{\frac{2k_BT_L}{\pi m_t^*}}\left(V_g - V_t\right)$$

Transverse effective mass continues to be reduced by strain hence drain saturation current will be enhanced even under ballistic transport condition which prevails under short channel condition. This will be more the case as the device is scaled. Thus strain will continue to scale well even in future generations as MOSFET goes deeper in Ballistic Mode of Transport.

50 years Journey of IC Technology_Appendix IX_History of SiGe HBT

Appendix IX gives the brief history of SiGe HBT and the role it is playing in revolutionizing personal communication.

Appendix IX.

History and Physics of SiGe HBT.

[Overview: Fabrication of SiGe HBT Bi-CMOS Technolgy- Cressler]

Table IX.1. History of SiGe HBT.

Device	Year of first intoduction	Reference
First epitaxial BJT	1960	Theurer,Kleimeck, Loar & Christemer, “Epitaxial Diffused Transistor”, Proceedings of IRE 48, 1642 1643,1960.

First SiGe HBT	1987	Iyer, Patton, Delage, Tiwari & Stork, “ Silicon-Germanium Base heterojunction bipolar transistors by MBE,” Technical Diget of IEEE International Election Device Meeting, San Francisco, 1987, 874-876.
First ideal SiGe HBT by CVD	1989	King, Hoyt, Gronet, Gibbons, Scott & Turner, “$Si/Si_{(1-x)}Ge_x$ heterojunction Transistor produced by limited reaction processing,” IEEE ED Letters, 10, 52-54,1989.
First SiGe HBT by UHV/CVD	1989	Patton, Harame, Stork, Meyerson, Scilla and Ganin, “ Graded-SiGe-Base, poly-emitter

		heterojunction bipolar transistors," IEEE ED Letters 10, 534 536, 1989.
Fiirst High performance SiGe HBT	1990	
First self-aligned SiGe HBT	1990	
First SiGe HBT ECL ring oscillator	1990	
First pnp SiGe HBT	1990	
First operation of SiGe HBTs at cryogenic temperature	1990	
First SiGe HBT BiCMOS Technology	1992	
LSI SiGe HBT IC(12 bit DAC 1.2GS/sec)	1993	
First SiGe HBT with peak $f_T = 100$GHz	1993	
First SiGe HBT in 200mm Wafer Technology	1994	

First SiGe HBT Technology optimized at 77K	1994
First Radiation Tolerance investigation of SiGe HBT	1995
First Report of Low Frequency Noise in SiGe HBT	1995
First SiGe:C HBT	1996
First High power SiGe HBT	1996
First sub-10psec SiGe HBT in ECL circuit	1997
First High Performance SiGe:C HBT Technology	1999
First SiGe HBT with peak f_T above 200GHz	2001
First SiG HBT with peak f_T above 300GHz	2002
First complementary symmetry amplifier using	2003

HBT	
First C:SiGe Technology with npn & pnp f_T above 100GHz	2003
First vertical SiGe HBT on thin film SOI(CMOS compatible)	2003
First SiGe HBT with both f_T and f_{max} above 300GHz	2004

Steps in SiGe HBT fabrication.

Step 1. Low Sheet Resistance Buried layer is fabricated: High dose Arsenic implant followed by a long thermal cycle in oxidizing ambient. This drives Arsenic deep in the wafer and anneals the defects caused by ion-implantation. A thin epi-layer is deposited at high temperature. The buried layer has a sheet resistance of $R_{sh} = 10\Omega/\square$.

Alternatively

Sub-collector may be just implanted in the wafer. This wafer is more compatible for CMOS processing. Burt since no following anneal step, hence to minimize the defects, dose and depth of implant is considerably less and $R_{sh} = 100\Omega/\square$.

Combination of deep trench isolation and buried sub-collector leads to reduced Collector Capacitance and reduced Collector Resistance therefore maximum **fT and fmax** can be realized.

In advanced HBT structure scaling involves both vertical and lateral scaling.

Vertical Scaling:

Vertical scaling consists of thinning all the three layers of HBT(base, collector and emitter). The base width is scaled by shrinking the base width and increasing Ge gradient across the base which includes increasing the peak Ge%. If all doped regions are thinned then transit time may decrease but RC time constant increases because R increases. To get an overall improvement in f_T overall time delay τ_{EC} must be brought down.

$$\frac{1}{2\pi f_T} = \tau_{EC} = \tau_E + \tau_C + \tau_B + \tau_{CSCL}$$

$$= \frac{kT}{qI_C} C_{EB} + \left(\frac{kT}{qI_C} + R_C + R_E\right) C_{CB} + \Upsilon\left(\frac{W_B^2}{D_n} + \frac{W_B}{V_{exit}}\right) + \frac{W_{CSCL}}{2v_{sat}}$$

Where

$$\Upsilon = \frac{kT}{\Delta E_{g,Ge(grad)}} \left\{1 - \frac{kT}{\Delta E_{g,Ge(grad)}} \left[1 - \exp\left(-\frac{kT}{\Delta E_{g,Ge(grad)}}\right)\right]\right\}$$

τ_E = RC delay time constant at EB junction;

τ_C = RC delay time constant at CB junction;

τ_B = Transit time delay through Base Width;

τ_{CSCL} = Transit time delay through reverse biased CB junction;

v_S = scatter limited velocity of electrons while falling down the potential hill at CB junction;

To achieve this objective , we must simultaneously reduce the two transit times, cross-sectional area of capacitances and series resistances. Introduction of C in SiGe base has helped reduce R_{sh} of Base even after thinning as required in vertical scaling.

Reduction in time delay requires that quiescent I_C is increased which means that Kirk Effect must be pushed to high Current Density. This is achieved by increasing Collector Dopant Concentration for which we do Selectively Implanted Collector(SIC).

SIC allows higher values of I_C, decreases R_C and reduces W_{CSCL}. But increase in SIC, means more lateral and vertical diffusion in subsequent heat cycles which leads to higher C_{CB}. So vertical and lateral scaling should be carefully controlled to maximize **fT and fmax.**

Table IX.2. Milestones of Development of SiGe-strained Si FETs.

Device	Year of first introduction
FET concept	1926
Si(MOSFET)	1960
Si(CMOS)	1963
First oxidation study of SiGe	1971
SiGe nMODFET	1986
SiGe pMODFET	1986
SiGe Photodetector	1986
SiGe SBD	1988
SiGe hole RTD	1988
SiGe BiCFET	1989
SiGe gate CMOS tech	1990
SiGe Wave Guide	1990
SiGe pMOSFET	1991
SiGe electron RTD	1991
SiGe LED	1991
SiGe Solar Cell	1992
SiGe photo transistor	1993
SiGe pMOSFET on SOI	1993
Strained Si pMOSFET	1993
Strained Si nMOSFET	1994
SiGe:C pMOSFET	1996
SiGe pFET on SOS	1997
Submicron SiMOSFET	1998

Vertical SiGe pFET	1998	
Strained SiCMOS	2002	

SiGe Technology has become the driving force behind the explosion in low cost, light weight, personal communication devices like digital wireless handsets and other entertainment and information technologies such as Digital Set-Top Boxes, Digital Broadcast Satellites, Automobile Collision Avoidance and Personal Digital Assistants. SiGe extends the life of wireless phone batteries and allows more durable communication devices. Products combining the capabilities of Cellular Phone, Global Positioning and Internet access in one package are designed using SiGe.

These multi-function, low cost mobile client devices capable of communication over voice and data networks represent a key element of the future of computing.

SiGe HBT in CMOS for high end PC failed but SiGe HBT succeeded in RF Communication Circuits because of low power consumption.

SiGe Technology is finding extensive applications in:

Wired Communication;

Wireless Communication Circuits;

In Disk Storage;

High speed BW instrumentation;

Discrete SiGe HBT in Amplifiers and Wireless Devices;

IC SiGe HBT in GSM handsets, in CDMA hand sets, in Base stations, in Wireless LAN chipsets and in high speed 10-40Gb/s Synchronous Optical Networks(SONETS) receivers.

Ge grading in Base helps higher transit frequency f_T and increase in short circuit current gain β_F . By suitable trading of β_F with r_x (base spreading resistance), BW can further be improved. By increased Base Doping, r_x reduces. This reduction adversely effects β_F but gives considerable improvement in transit frequency and in Noise Figure of the device. For same I_C (collector current) , SiGe HBT has a higher short circuit current gain, lower RF noise and lower flicker noise or 1/f noise as compared to an identical Si BJT. In SiGe , higher raw speed and lower power consumption can be traded depending upon the application.

Table IX.3 Comparison of CMOS with conventional Si-BJT and SiGe HBT (after Harame)

Parameters	CMOS	Si BJT	SiGe HBT
f_T	high	high	Higher
f_{max}	high	high	Higher
Linearity	Best	Good	Better
V_{BE} (or V_{Th}) tracking	Poor	Good	Good
1/f noise	Poor	Good	Good
Broad Band Noise	Poor	Good	Good
Early Voltage	Poor	OK	Good
g_m	Poor	Good	Good

Real strength of SiGe lies in Analog, RF and Digital Applications in existing CMOS Fabrication Foundaries. This makes possible implementation of new Architecture such as direct conversions and software defined radio.

SiGe BiCMOS Technology.

This technology demands that Low Temperature Epitaxy(LTE) SiGe process be combined with high temperature CMOS processing.

CMOS performance must be retained (same as the parent CMOS process) after the addition of LTE SiGe in order to use existing digital ASIC Libraries and Design Methodologies.

Similarly CMOS processing must not significantly alter the doping profile (and hence performance) of SiGe HBT.

In this integration the two primary issues are thermal budget and trade off between process modularity and process sharing.

IBM first generation (5HP) 0.5μm SiGe BiCMOS used a “base equal gate” scheme. A common layer stack is used for both HBT base and FET poly-Si Gates.

Problem arises when CMOS advances to 0.24μm Technology(6HP). At this point CMOS thermal cycle increased significantly because of the need for Source/Drain dopant activation for NMOS and gate side wall oxidation. So “ base after gate” strategy is adopted. HBT is built after the formation of gate, gate spacer, LDD implants and NMOS anneals. This simplified BiCMOS integration of SiGe HBT with newer generation of CMOS.

Solid State Physics & Devices _The harbinger of third wave of civilization_PREFACE_Unfolding of Third Information Revolution.

Summary: In the first part the lecture gives the background of Third Information Revolution. Then it describes the unfolding of Third Information Revolution. It traces the development in Electronics in general and in Computer Engineering in particular from 1833AD to date. It gives the development of SuperComputers till date. It gives the source of digital data and total data generated per annum. It gives a Table on World Wide Internet Users. Table 1 gives the evolution of μP chips, Table 2 gives the evolution of different generations of computers, Table 3 gives the evolution of Local Area Network, Table 4 gives the data on Submarine and land cables laid till now. It also gives a special note on Hard 350 Disc Drive introduced in IBM 305 Ramac Computers supplanting the magnetic tapes.

ROAD MAP OF COMPUTER DEVELOPMENT

[Computer age- Communication fuelled Universal Empowerment]

Key words: Telegraphy, Vacuum Tube, Transistors, integrated Circuit Technology, Computer;

Summary: In the first part the lecture gives the background of Third Information Revolution. Then it describes the unfolding of Third Information

Revolution. It traces the development in Electronics in general and in Computer Engineering in particular from 1833AD to date. It gives the development of SuperComputers till date. It gives the source of digital data and total data generated per annum. It gives a Table on World Wide Internet Users. Table 1 gives the evolution of μP chips, Table 2 gives the evolution of different generations of computers, Table 3 gives the evolution of Local Area Network, Table 4 gives the data on Submarine and land cables laid till now. It also gives a special note on Hard 350 Disc Drive introduced in IBM 305 Ramac Computers supplanting the magnetic tapes.

[**Science &technology year by year** –Marshall Publication,London,2001]

Gutenburg Revolution, the Second Information Revolution , gave rise to the rotating type Printing Press. "Printing Liberated people's mind by enabling the widespread dissemination of knowledge and opinion. But it remained subject to censorship and worse. Television shrank the planets into Marshall Mclean's "global village" but the representation it diffused were tightly controlled by editors, corporations, advertisers and governments."

The net in contrast provides the first totally unrestricted , totally uncensored communication system-ever. It is a living embodiment of an open market in ideas. Its patron saint (if that is not a blasphemous though for *such a secular hero) is Thomas Paine, the great 18 th century libertarian who advanced the remarkable idea (remarkable for the time, anyway)that everyone had a right to speak his mind and the even more notion that they had a right to be heard.”*

Precisely because of this “Net is potentially more powerful than both printing and TV because Net harnesses the intellectual leverage which printing gave to the man-kind without being hobbled by the one to many broadcast television.”

BC700- *The First Information Revolution* - Invention of the Phoenicians Alphabets, the mother of all European languages.

BC 400 –Invention of Brahmilipi , the mother of all Indian language except Urdu.

100 AD –Invention of Kharosti language and alphabets, the mother of all Arabic

languages including Urdu

1450AD –*The Second Information revolution* – Invention of rotating press by a German Mechanic named Gutenburg.

1833AD – Michael Faraday made the earliest study of semiconductors.

1834AD –Charles Babbage plans a machine to calculate and print results.

1842AD –Ada Lovelace writes a step by step Programme for Babbage’s Machine.

1850AD –George Boole invented **Boolean Algebra.** Boolean Algebra is the basis of the

present set of Computer Systems which use **ALGORITHMIC**

PROGRAMMING as opposed to **HEURISTIC PROGAMMING** of

Artificial Intillegence Computers. Heuristics programming uses **Predicate**

Logic. Heuristics means working by rule of thumbs and it is self learning

programme. In contrast Algorithm means a set of sequential steps which are

repeated or iterated. Here there is no self learning involved.

1851AD-First successful commercial telegraph cable laid from England to France. The cable had a central core of copper wire, gutta-percha, tarred yarn, tarred hemp, and iron wires.

1858AD-First transatlantic telegraph cable.

1876AD – Alexander Graham Bell invented Telephone.

1895AD –J.J Thomson discovers electron as the basic building block of Atom through

the study of Cathode Ray;

1902-1906AD- Telegraph cable crosses the Pacific ocean.

1904AD-John Ambrose Fleming invents Vacuum Tube Diode(two electrodes device).

1906AD-Lee De Forest invented Vacuum Triode (three electrodes device) and used it to

achieve RC coupled Amplifier. Subsequently four electrodes device tetrode and

five electrodes device pentode.

1911AD-Using Heaviside "Distortion-less Transmission Condition" i.e. loading the

Telephone cables with large inductors, first long distance telephone call was

made from New York to Denver, Colorado, over a distance of 2000 miles.

1913AD-Vacuum Tube Repeater introduced in

Public Service Telephone

Network (PSTN).

1915AD-Using inductor loading & R.C Coupled Amplifier, first transcontinental

telephone call is made from New York to San Francisco across a distance of

4000miles. Telephone becomes a Public Utility Service.

1918AD-Two Frenchmen, Block & Abraham, build the first simple Calculator-a

calculator with programme storage and data storage facility is a computer.

1937AD-Alan Turing invents “The Turing Machine “- it can perform all logical tasks.

1943AD- To decipher encrypted messages, British Scientist build “Colossus”-

forerunners of Electronic Computers.

1945AD-IBM builds first full scale calculator, MARK-I.

1945AD-ENIAC (Electronic Numerical Integrator & Calculator) build by Moore School of Electrical Engineering at the University of Pennsylvania. ENIAC is 27Metric Tonnes , consumes 200kW and occupies a large room. The valves failed at a rate of one every 10 minutes and ENIAC had 18,000 valves. So development of user friendly home Computers was a pipe dream at that time. Development could not take place using Thermoionic Valves.

1945AD- Microwave Telephone link established.

1945AD-First Transoceanic Telephone Links established using voice quality coaxial

cables. Telegraphic Cables are not suitable for Telephone Conservation. Voice

quality cables were developed during World War II and introduced right after

the War .Before the World War II , no TransAtlantic or TransPacific Telephonic

Links were there. Across the oceans only Radio-Telephone call could be made.

1947AD-At Bell Laboratories, John Bardeen, Walter Brattain & William Shockley

invented BJT (Bipolar Junction Transistor). They were awarded Nobel Prize in

1956 in Physics for this invention.

1948AD-Magnetic Drum Memory invented at Georgia Institute of Technology for

Computers.

1952AD-UNIVAC I (Universal Automatic Computer)- First computer designed for

business use. The inventors were John Mauchy & John Eckart.

7.26 Metric Tonnes, 5000 Vacuum tubes & 1000 calculations per second. These are termed **FIRST GENERATION COMPUTERS**.

1955AD-Semi Automatic Ground Environment (SAGE) installed. This takes in data

received by the RADAR network and computes interception trajectories and

steers the defenders within 1000yards of the attacker.

1956AD-UNIVAC II using transistors come in the market. Manufactures are Sperry

Rand, USA. These are termed as **SECOND GENRATION COMPUTERS**.

1956AD-TAT-1, the first transatlantic voice grade telephony coaxial cable laid across the Atlantic Ocean.Until then transatlantic telephone calls were made by Radio-Telephony.(Coaxial Cable, polythene insulation, vacuum tube repeaters).

1956AD- September 13, 1956, IBM 305 Ramac Computer with DISK DRIVE was introduced. Till then the OS and data were stored on Magnetic Drums or Magnetic Tapes as a result the booting process of the computer or access of data was very slow. This newly introduced computer had 350 Disk File. This consisted of a rack of 24” magnetically coated platters 50 in number mounted on a single vertical spindle and could store 4.4MB of data as compared to today’s hard disc which store 70GB of data.[Please see the foot note].

1959AD-Integrated Circuit (IC) chip invented by Jack Kilby (Texas Instrument) and

Robert Noyce (FairChild). Discrete active and passive devices are integrated

together on one wafer thin silicon chip in the area of 1cm by 1cm square,

mounted on a header and then properly encapsulated and sealed. At the time it

started, it had only 30 components in the 1 square cm area. This heralded the Information Age and **THIRD Information Revolution**. It also produced an

Explosion in electronic consumer and communication applications such as Personal Computers to cell Phones. Jack Kilby was awarded the Nobel Prize in physics in the year 2000. Jack Kilby passed away on 1st August 2005. Robert Noyce went on to establish “ INTEL” (Integrated Electronics) along with Gordon Moore and Andy Grove . He passed away earlier than Jack Kilby so he could not be awarded. Noble Prizes are not awarded posthumously.

5th October,1959-IBM1401 mainframe computer makes its debut. This becomes the best selling computer in the world in mid-1960s.This was in the category of stored program computers. These stored program computers were vastly superior to IBM 604 Electronic Calculating Punch Machine which was vacuum tube, plugboard-controlled, serial-decimal machine with 50 digits of storage.Stored Program Computers were more flexible and adaptable than plugboard-based accounting machines. But these large-scale stored program computers were too expensive for common use.By the use of alloyed junction discrete transistor logic and Automatic Logic Design software, the volume and power consumption was reduced by 50% and 95% respectively.Standard Modular System(SMS) called Cube/Rolygon for electronic packaging was used.

Saturating Complementary Transistor Diode Logic(CTDL)Family was used. This could clock at a maximum rate of 250kHz.CTDL was robust and handled large fan-ins and fan-outs.Memory to Memory architecture was used.This reduced the number of instructions needed for accounting and business programming.These changes achieved a entry-level rental price target of US$2500 per month.Previously it was as high as US$40,000 per month. It had outstanding print quality, powerful magnetic tape subsystem and a stored program computer for the mass market place.By 1965, half of the approximately 26000 computers in the World were 1400-family machines.Tape-oriented 1401 system became the computer of choice. Time-life transferred 40 million punched card subscriber records to just several hundred magnetic tapes.Full size system had 500,000 discrete components, weighed up to 4 tonnes, and consumed up to 13000W.It used 10,600 Ge alloy-junction transistors and 13,200 Ge point-contact diodes on 2300SMS cards interconnected with 5.5mi of wire. Memory varied from 1400 bytes to 16000 bytes memory space.It had a clock rate of 87000cycles/sec corresponding to one clock cycle of 11.7microsecond. In 50 machine cycles corresponding to 0.5ms two positive 20-digit numbers could be added. Today a Pentium IV processor operating at 4GHz can add two 64-bit numbers in 0.5nano second.Magnetic Core Memories were priced at 60 cents per bit(or US$24 per byte in today's currency. This is 300 million times more costly than today's DRAM.Its high speed and enduring print quality made it an Industry Workhorse. By 1961, optical character recognition (OCR) devices were introduced as a peripheral to 1401. The 1401 contributed to the growth of programming as a profession and software as an industry.It had no operating system. IBM had a compatibility problem amongst its various versions and amongst its various peripherals. By April 1964 it introduced System/360. It consolidated software, peripherals and support in one compatible and scalable computer family. 1401 wound down with the announcement of S/360.[This part has been excerpted from"The Legendary IBM 1401 Data Processing System" by Robert Garner Fredrick(Rick) Dill, IEEE Solid-State Circuits Magazine,Winter 2010,Vol.2,No.1, pp.28-39]

1960AD-PDP I (Programmed Data Processor).This was introduced by Digital Equipment

(DEC).DEC was founded by Kenneth Olsen & Harlen Anderson .These were

also made of transistor. These were used as Office Computers. They were of

cabinet size. Control Panel and the Key Board sat on the desk. It cost $

120,000.00. This was way too expensive for most customers and for most

application. This started the era of Mini-Computer.

1960AD-Man-Computer Symbiosis was published by J.C. Licklider, an experimental

psychologist. This article stated that rote algorithm for computers & creative

heuristics for Humans put together could be far more powerful than either could

be separately.

1960AD-Sir Arthur's Clark's dream of Global Satellite Communication is realized.

1961AD-FIRST RLT (Resistance Transistor logic) IC Chip is commercially marketed by

T I (Texas Instrument) and FairChild.

1964AD-Semi-Automatic Business- Related Environment (SIBRE) is created by the

introduction of Nation-Wide Ticketing System like our Nation-Wide

Computerized Reservation System today in 2005 in Indian Railways.

1964AD-FIRST Analog IC Amp. Chip MC 1530 marketed by Motorola. Op. Amp.

Stands for Operational Amplifier which were in yesteryears used in Analog

Computers for different Mathematical Operations and hence the name Op.

Amp.

1964AD-PDP 8, table top computer introduced in the market .It was a 8-bit computer

meaning by the address size is 8 bit. Larger is the address word, larger is the

address capability. 4-bit means 16 locations can be addressed and 8-bit means

256 locations can be addressed. This used IC Chips hence it was compact and

less power consuming. This was **THE THIRD GENERATION**

COMPUTER.

Price $18,000.00, 125 kg, and 4kBytes memory.

1965AD- Project MAC, started by Licklider at MIT, evolved into first On-Line

Community, complete with bulletin boards, e-mails , virtual friendship and a

“free ware” exchange.

1968AD- MODEMS are used for Wide Area Network (WAN).

1968AD- December1968, demonstration at Fall Joint Computer Conference in San

Francisco established Graphical User Interface(GUI) with mouse, on screen

windows, full screen word processing and a host of other innovations.

1968AD-18th July, INTEL (Integrated Electronics) was established by Robert Noyce and

Gordon Moore . Later this team was joined by Andy Grove.

1970AD-Advance Research Program Agency (ARPA) established ARPANET, the first

WAN, connecting Stanford Research Institute, University of Utah, University of

California Los Angels and University of California Santa Barbara mainly for

Defence research purposes. In three years it grew to cover whole USA.

1970AD- PDP-11 introduced with 16 address word

1970AD-Magnetic Disks replace Magnetic Drum as back up memory. These are called

Floppies (8” or 20cm) or Diskettes(3.5”or 9cm).Memory capacity is 1.44MB.

1971AD- FIRST INTEL MICROPRPCESSOR CHIP (μP 4004) introduced in the

market . It can handle 4-bit data word and 8-bit address word meaning by it can

address 256 locations. This 1 chip combined the function of 12 subsystems of a

CPU(Central Processor Unit). This one silicon chip (1cm × 1cm)had 2300

transistor , system clock rate 108kHz, Cost $299.00 . In India it was available

for Rs.7500.00.

[A development contract from a Japanese Company , for a set of chips needed to power an electronic calculator , triggered off what became the world’s **first computer-on-a-chip: the Microprocessor.** 35

years ago, INTEL, a small Santa Clara (US) based manufacturer of memory and switching devices, unveiled a thumbnail-sized slab of silicon encased in a ceramic casing with 16 pins: this ushered in the era of Personal Computers, the era of Desk Top Computing and the era of **FOURTH GENERATION COMPUTERS.**

The contract with Busicom required INTEL to deliver the electronics for calculating machine, as a set of 12 custom built chips. The electrical engineer assigned to handle the task-Dr. Marcian “TED” Hoff , a PhD from the Stanford university-had an inspired thought. Why make a set of chips just for one application? Why not generalize the design so that other computing tools could be built with the same generalized design?

In the process ,Dr. Hoff also suggested that the composite elements of a computer , as it was then understood- a unit to do the arithmetic (ALU-Arithmetic-Logic-Unit), a small memory and input and output interface –could be combined in a single slap of silicon..

When Busicom had cash flow problems, INTEL bought back the design it had created and slipped the product into its own catalogue.

Intel µP 4004 was sold to Defence Establishment in India at a cost of Rs 7,000.00. Intel’s Indian Agent was Hyderabad based Electronics International which was later named as Microelectronics International]

1971AD-Texas Instruments marketed first pocket calculator 1kg in weight and price $150.00. It was too bulky.

1971-Electronic Mail or e-mail is invented by Ray Tomilinson.

- @ is introduced by Tomilinson to separate the User Name from domain name.
- Tomilison combined the existing mail program ‘SENDMSG’ that worked only within an organization with a file transfer program called ‘CPYNET’ to create an e-mail public utility between 15 computers within the organization and use the same on the existing ARPANET.
- The first e-mail; was “QWERTYUIOP”, the second row of the typewriter keyboard, sent by Tomilison.
- The first telegraphic message was sent by Samuel Morse in 1844. It was “What hath God wrought”.
- The first conversation carried out between Alexander Graham Bell and his assistant Watson in 1876 was, “Mr. Watson comes here. I want you.”.
- Founder of Hotmail got the idea of accessing email via the web from a computer anywhere

in the world. When Sabeer Bahtia came up with the business plan for the mail service , he tried all kinds of names ending in ‘mail’ and finally settled for Hotmail as it included the letters “html’(hyper text mark up language), the programming language used to write web pages .It was finally referred to as HoTMaiL with selective upper casings.

1972AD-first 8 bit μP 8008 introduced by intel.

1973AD-May 22, ETHERNET IEEE 802.3standards for LOCAL AREA NETWORK were born.

The inventor of Ethernet was a young Xerox Palo Alto Research Scientist named

Robert M. Metcalfe.

1974AD-8080μP chip introduced.

1974AD- First paper on TCP (Transmission Control Protocol)and the architecture of the Internet

by Vinton Cerf and Kahn.

1975AD-Altair 8800(mini-computer architecture) based on Intel μP Chip 8080 (8 bit

data & 16 bit address word) introduced in the market. $397.00 worth kit available.

1976AD-APPLE CO. founded by Steve Wozniak & Steve Tabs members of Home-Brew

Computer Club (California) in a Garage. Apple I was introduced.

1976AD-Cray 1 designed by Seymour Grey. This was a class apart called Super

Computer. It cost $8.8million. It could achieve 240MFLOPS (Floating Point

Operations per second)

1977AD-Taito (Japanese Software Company) introduced a video game “SPACE

INVADERS” which takes the world by storm. The first video game was introduced in 1970’s by ATARI. It was a ping-pong game.

1977AD-First Personal Computer was introduced, named APPLE II , at a price of

$1300.00 .It had 4kB RAM. It had a software VisiCalc the best spread sheet

package of its day. This was formally the **FOURTH GENERATION**

COMPUTER.

1977AD-Bill Gates, Harvard Under-Graduate dropout, and Paul Allen developed the

programming language BASIC for Altair 8800 and in the process founded

MICROSOFT Company at Albuquerque near MIT. The name was coined by

Bill Gates to represent the company that was devoted to MICROcomputer

SOFTware. MS provided DOS (disc operating system) to IBM and other PC

makers.

1978AD-Intel introduces 8086 μP Chip which had 16bit data word and 16bit address

word.

1978AD-HCL delivered 8CR μP based Desk Top computer.

Indigenous OS;

Indigenous BASIC Interpreter Languages.

1979AD__On July 1, 1979, Sony launched Walkman which played a magnetic tape popularly known as cassette but there was no provision for recording. While walking or moving, you could play a cassette and listen to it. It became so popular that within next two months , thirty thousand pieces were sold. To this date on 27.10.2010 altogether 385 million pieces have been sold. The production by Sony has been stopped as Apple-iPod has out sold Walkman. Apple-iPod has solid state memories and digital signal processing. Only the final stage where the music has to be played is an analog voltage amplifier. All these combined makes Apple-iPod much more handy, economical and much more diverse interms of variety of music which can be played. Hence Apple-iPod has caught the fancy of young people.

1980AD-Voice to text system was developed by IBM.

1981AD-IBM introduced its own brand of Personal Computer PC5150 based on 8088μP

Intel Chip. Clock Rate was the same as that of APPLE II but memory was much

bigger at 40kB. GREEN SCREEN, 2x5” Floppy Drives for loading OS. IBM

PC design accounts for 90% PCs. APPLE PC design accounts for 10% PCs.

1982AD-France Telecom launches MINITEL system, a wide area network in which you

can login to a central data base from any where in the country and

pertinent information can be availed. This is first Public Utility On-Line

Service.

1982AD-Intel introduces 80826μp Chip (16 bit data bus and 24 bit address bus and 68

pin package).

1983AD-MOUSE introduced in LISA, an Apple machine. It had a user friendly

graphical user interface (GUI). Its patent was bought from Xerox research center where it was used in Xerox Star. Screen icons were introduced which could be selected by the use of cursor using a mouse.

1983AD-CD_ROM (COMPACT DISC_READ ONLY MEMORY) introduced. It has

700MB as opposed as 1.44MB in floppies. This can accommodate one hour of

music .

1983AD: TCP/IP, standardized in 1978, rolls out formally 1 January 1983 and all ARPA-sponsored packet networks (ARPANET, Packet Radio Net, Packet Satellite Net) and, subsequently ethernets that are becoming commercially available, incorporate the standardized version of TCP/IP.

1984AD-APPLE MACINTOSH is introduced. It is a complete stand alone system with

the latest GUI, namely mouse and screen icons. Its cost was $2500.00.

1984AD-Musical instruments digital interface (MIDI) introduced.

1985AD-Desktop publishing introduced by ALDUS PAGE MAKER along with LASER

JET PRINTER in Apple Macintosh.

1986AD-80386(Intel) and 68030(Motorola) introduces 24 bit address bus.

1988AD-VIRTUAL REALITY (VR) is introduced. When VR headset is put on , you see

a world of your choice with the help of computer. You can place yourself in the

cockpit of a jet plane and experience the sensation of actual flight.

1988AD-Pattern Recognition Software introduced to read hand written documents by

computer .

1988AD-MPEG (Moving Picture Expert Group) is devised for compressing digital

videos Band Width.

1988AD-First Trans Atlantic Optical Fiber Cable, TAT8, used in place of coaxial cable.

Coaxial Cable can carry 3,000 simultaneous phone calls. In contrast Optical Fiber cable can carry 37,500 simultaneous telephone calls with the present state of art of electronics. As the switching speed increases, the number of simultaneous calls increases exponentially. Total BW: 5.12terabits/sec.

1988AD-Touch Screen Computers are developed.

1989AD-Disk Operating System (DOS) are replaced by MS WINDOW OS., MS EXCEL

and MS WORD introduced. WINDOW 3.0 VERSION released.

1989AD- March 1989, computer software child prodigy, now an adult, Mr. Tim Berners-Lee of CERN, Geneva, handed a proposal ‘Information Management: a proposal’. They came up with global hypertext language, the basis of “http” in website address. By October 1990, they developed the first web browser. The World Wide Web Technology was made available for wider use on INTERNET from 1991 onward without any royalties. “Internet is a vast network of networks, interconnected in many different ways yet they all speak the same language. Web is one-albeit the most influential and well known-of many different applications which run over the Internet”. Today in 2009 March, as we celebrate 20th Anniversary of Web Technology, Tim Berners-Lee is a researcher at MIT,US, and a Professor at Computer Science Department at South Hampton University, Britain, and still heads the World Wide Web Consortium that coordinates development of the Web.

1991AD-ARPANET was confined to exclusive clientele (DOD sponsored research). To

make the NET available to University Community at large, NSF(National

Science Foundation) connected his own NSFNET to ARPANET . This led to

exponential growth in the use of NET. In mid-90’s the collection of networks

was looked at as INTERNET.

1991AD-Until 1991, internet was largely populated by academic, government and

industrial researchers. Tim Bernes Lee developed Hyper Text Mark Up

Language (HTML) for easy file transfer and from here he was able to build up

World Wide Web (WWW). HTML is the lingua franca of WEB pages. This

made Internet useful for the whole populations – academic and non academic

both.

1992AD-Mobile Phones enter Digital Age. These are second generation (2G) mobiles.

1993 AD-PENTIUM I introduced . it had 3.1x10^{6} CMOS(Complementary Metal Oxide

Semiconductor Field Effect Transistor) placed on one chip .

- it could perform 90MFLOPS, 5 IMES THE SPEED OF 80486.
- data bus 32 bits and address bus 32 bits
- clock rate 1500 times that of 4004 .

1995AD-WINDOW 95 was introduced .**1996AD-**Multi-media Extension Pentium Chips (MMX) Chips launched. It had $5.5x10^{6}$

transistors (CMOS). It can perform 300MFLOPS and speed up multi media

processes.

1998AD-APPLE launches iMAC fully internet compatible.

1998AD-Global Positioning System (GPS) introduced by US Defence.

1999AD-IBM’s DEEP BLUE defeats Garry Kasporov (chess grand master). 200 million

different chess positions are examined per second.

1999AD-DVD(Digital Versatile Device)introduced with a memory space of 8 GB as

compared to 700MB in CD. DVD became a reality only after the invention of

Blue Laser Diode .

1999AD-HDTV(High Definition TV) introduced.

1999AD-WEB CAM (WEB camera). Digital Camera connected to your computer. It

has a special drive which allows a picture to be taken every 10 seconds, turns

this into a standard file format and uploads to the server where your website is

stored . People can keep visiting the site and have a look at the photographs of the location as desired.

- Handheld Internet Access introduced;

- Hand held Personal Organizer introduced;

- MP3 standard is introduced for Digital Audio Compression.

2000AD- Blue Tooth Technology introduced . This Wireless Technology allows computers, phone hand sets and CD players to communicate and share data over a distance of 10m and less. Operating frequency 2.45GHz. Switches the operating frequency 1600 times per second . This is called Frequency Hopping Spread Spectrum method. When Blue Tooth capable devices come within range of one another they have electronic conversations.

2001AD-Pentium IV chip introduced;

- Nano-electronics chip;

- It has 42 million transistors packed in an area of 1 cm by 1 cm;
- System Clock Rate is 2GHz;

- 20,000 more powerful than 4004 μP Chip;
- Cost Rs 15,000.00

2001AD- 23rd October 2001, Apple iPod, the iconic device that defines our era as distinctively as the Sony Walkman defined 1980s. An entire ecosystem of goods and services have evolved around Apple iPod. Minispeakers can be plugged into it; microphones attachment can convert it into digital audio recorder; it can be turned into small radio transmitters that beam songs to the nearest FM radio; attachments can turn it into breadth analyzer. iPod plus iTunes turned out to be a grand success where other musical gadgets and soft-wares failed.

2003AD-APPLE introduces Power Mac G5 PC.

- 64 bit data word and 64 bit address word;
- 18 PETA Memory locations can be addressed;
- System Bus operates at 800 MHz;
- System Clock operates at 2GHz, dual processors;
- 80 to 160GB hard disc;
- 256MB to 512MB high speed RAM;

2004AD-Internet Protocol Version 6 (IPv6) introduced. Till now we have been using

IPv4;

- Address field is only 32 bit long in IPv4 but 128bits long in IPv6. This will allow much larger number of users than what it can support till now;

- New version allows a larger through-put;
- New version offers a better support for options;
- New version provides better methods of authentication and privacy. This will

provide a much better protections against Hackers;

2004AD- " Project Columbia" a $50 million super-computer built by SILICON

GRAPHICS is introduced;

- it is powered by 10,240 processors in 20 units;
- it is Intel Itanium chips;
- it performs 43 Tera FLOPS as compared to the world record of 36TFLOPS

performed by IBM's BlueGene/L super computer;

- This is a ten fold increase in NASA's supercomputing capacity;
- Applications in hurricane tracking, weather modeling and earth's interior

imaging;

2004AD-Norton Internet Security 2005 is introduced for Rs 3,618.00. It offers Norton

Anti-Virus, Norton Anti-Spam and Norton Personal Firewall. Firewall protects

the system from Trojan Horses and other attackers. It also includes parental and

privacy controls;

2004AD- Radio frequency Identification (RFID). It is a stamp size chip which gives

a Radio Signal as the tag mark of the product it is attached to;

2005AD- GeForce 7800 GTX (latest Graphics Processing Unit or GPU) introduced by

US-based nVIDIA Corporation for Video Games;

- 302 million CMOS on one 1cmx1cm chip;
- Performs 202 GFLOPS;
- 10% less power consumption ;
- 50% quieter
- twice as powerful as its processors 6800 but it costs the same as before that is

Rs. 32,000.00 ;

- Market for video games is one of the fastest growing sectors in Computer

Industry;

2005AD-Dual cores Microprocessor chip are introduced;

Intel introduces Pentium “Extreme Edition” 840 with Clock Rate = 3.2GHz and

with billon transistors . This is dual core as well as it has a software called

‘hyperthreading’(HT). This is a software which enables one to work with double

virtual processors. Thus a dual core with HT will allow application developers

to write software that can carve up the task between four processors;

- Intel introduces Xeon for servers and Pentium M for laptops;
- AMD introduces Opteron 800 series for servers and workstations;
- AMD introduces Athlon 64 X2 for consumers and business PC’s;

2008 October. GeForce 9400M is a single chip solution for graphical processing platform for portable applications. It has computing rate of 55GFlops (Floating Point Operations), the most powerful integrated graphics processor in the market in 2008. It has 6 cores operating in parallel. But much less power hungry. New MacBook family will be fuelled by these chips.

2008 November. Intel and AMD have launched a native quad processor unlike the earlier version where 2 dual core chips had been combined in the same package. Intel has launched ‘Nehalem’,the i7 processor using 45nm technology. AMD has launched 45nmOpteron. Both achieve the 4 cores on the same silicon slab. This has Memory Bandwidth. These conserve energy by cutting down power when the device is idle.

2009 May- Interplanetary Internet Protocols are tested by using computers on the International Space Station. Eventually there will be an internet connectivity among Earth, multiple probes, rovers, orbiters and space crafts exploring the solar system. This will be the basis of Interplanetary Internet. The network’s key technologies are called Delay Tolerant Networking(DTN) protocols. The problem in deep space internet is the huge delay and occultation of the nodes. To overcome this problem “store and forward” technique will be used. The node will save the data and then passes on when the link is again established. A third test is yet to be carried. In this a security protocol and a new file-transfer protocol will be used.

2010AD-Pacific Unity will lay multiple fibers. Total capacity; 7.68 terabits per second.

ROAD MAP OF SUPERCOMPUTER.[‘The Data’, IEEE SPECTRUM, APRIL 2009]

1976 Illiac IV, 64 processors, teraflops capability, cost was 4.8×10^{-7}TeraFlops per million dollars or 4.8×10^{5}Floating Point Operations per Second which is 480×10^{3}FLOPS or 480kFLOPS per million dollars.

1976 CRAY-1, 1 processor with teraflops capability, cost was 1.778×10^{-5}TeraFlops per million dollars or 17.78MFLOPS per million dollars.

1988 CRAY Y-MP, 8-vector processors with teraflops capability, cost was 1.115×10^{-4} TeraFlops per million dollars that is 111.5MFLOPS per million dollars.

1997 ASCI RED, 4510 processors with teraflops capaability, cost was 1.818182×10^{-2} TeraFlops per million dollars which comes to 18.18182GFLOPS per million dollars.

2002 EARTHSIMULATOR , 5120 processors, cost was 17.5GFLOPS per million dollars.

2004 BLUEGENE/L, 65536 processors with 10TFLOPS capability, cost was 2.8TFLOPS per million dollars.

2007 PLAYSTATION3 CLUSTER, 8playstation 3s, 375TFLOPS per million dollars†.

2008 NVIDIA TESLA, 960 cores, 439.1 TFLOPS per million dollars*.

2008 ROADRUNNER, 19440 processors with several tens of TFLOPS capability, cost was 8.3TFLOPS per million dollars‡.

*Based on the cost of building your own Tesla supercomputer out of four Tesla C1060 units. Complete instructions are available at nvidia.com

† Seven processors per node at 150GFLPS per Unit for a single processor(11 processors per node for a dual processor). $400-plus per node.

‡ 6480 Opteron CPUs and 12960 Cell processors.

Supercomputers are massively parallel chips can be used to calculate protein folding, predict climate change and crack the encryption of hitherto-secure Web sites.

Nvidia advertises its new workstation, the Tesla, as a “personal supercomputer”. It clusters 4 Nvidia C1060 processing boards, each of which unites 240 graphic cores to process instructions at nearly teraflop speed.

Tesla does single-precision floating-point calculations using 8-bit bytes.

Roadrunner uses 64-bit floating integers.

Digital Data Generation.

Sources of digital data:

1. social networking sites,
2. internet-enabled mobile phones and
3. government surveillance.
4. 70% of the digital content is created by by individuals through phne cals, emails, photos, online banking transactions or postings on social networking sites.

Digital Universe Report in 2007 gave total digital content as 161 bn gigabytes.

Digital Universe Report in 2009 gave total digital content as 487 bn gigabytes.

IDC/EMC estimate:

The cost of the computers, networks and storage facilities that drive the digital universe is $6 tn.

Medical equipment, entertainment and content creation equipments cost another $6 tn.

Total World GDP is $50tn.

80% of PC's are powered by Intel Chips and remaining 20% powered by Motorola, Digital Co., AMD and Transmeta.

2006AD-We have reached 1 billion mark in personal Computers. 39% is accounted by USA, 25% is accounted by Europe and 12% by Asia.

PC user per hundred of the population- (USA 70%), (France 35%), (Brazil 7%), (China 3%), (India 15 per 1000);

2006AD-2 billion Cellular Phones- worldwide;

China- 450m cell phones; India-95m. cell phones

World Wide Internet Usage [www.internetworld-stats.com] Total User 1.3 billion

Countries	Total Number till March 2006	% penetration in the total population
India	30m	5.4
China	123m	9.3
Japan	86m	67.2
USA	207m	69.3

To bridge the Digital Divide, Government of India is planning to set up Common Service Centers (CSC)-broad band enabled computer kiosks that will offer a range of government-to-citizen and business-to-customer services, besides providing sheer access to the Internet. About 100,000 plus CSC's will be introduced in 600,000 villages by March 2008.

Companies in PC manufacturing – WIPRO, PSI, ZENITH, DCM,HCL

-27% growth rate -5m. unit/yr

Table 1.Evolution of Microprocessor Chip

Name	Date & Address Size	Packing Density	Number of Pins	Clock Rate	Instruction Per Second	Year & Price.
Intel 4004	4-bit	2300 PMOS	16	108kHz	?	1971 Rs 7,500.00
Intel 8080	8-bit	6000 NMOS	40	2MHz	0.64MIPS	1977
Intel 8086	16-bit	29,000 HMOS	40	5 MHz	0.33MIPS	1978
Intel 80286	Data 16b Address 24b	134,000 HMOS	68	6-12.5 MHz	1 MIPS	1982
Intel 80386	32-bit	275,000 CMOS	132	16 MHZ	5 MIPS	1985

Intel 80486	32-bit	1.2million CMOS	168	25 MHZ	20 MIPS	1989
Pentium	Data 64b Address 32b Memory Size 2_{32}–4GB	3.1 million CMOS	?	60 MHZ	100 MIPS	1993
Pentium II	Same	7.5 million CMOS	370	233 MHZ	400 MIPS	1997
Pentium III	same	9.5 CMOS	370	450 MHZ	1000 MIPS	1999
Pentium IV	Data 64b Address 64b	42 million CMOS(Nano electronics)	478	2.2 GHz	?	2001

MOORE's LAW: The packing density of CMOS on µP Chip would double every two years leading to an exponential growth in the complexity and speed of µP Chip.

Table 2. Evolution of Computers.

Generation	Components	Architecture	Logic	Programming	Language
	Von-	Boolean	Algorithm	Fortran,Basic,Algol	

		neumann, sequential				
5th	Robotics, Data Based Medical Diagnostic Tools	Artificial Intelligence Machines	Data-flow, Systolic, Pipeline	Predicate	Heuristic	Prolog, Lisp

Table 3. Evolution of Local Area Network

Year	Standard	Name	Data transfer rate
1973	802.1	OSI (Open System Interconnection)	10Mbps
	802.2	Logical Link Control	10Mbps
1983	802.3	CSMA/CD (Carrier Sense Multiple Access/ Collision	10Mbps

		Detect)	
	802.4	Token Bus	100Mbps
	802.5	Token Ring	100Mbps
1999	802.11	WLL (Wire Local Loop)	1 Gbps
2001	802.13	Wi-Fi (Wireless internet Access	

Table 4. Development in submarine cable. (Sources: Telegraphy Research: *"History of the Atlantic Cable and Undersea Communications"* **by Bill Glover,Wikipedia.**

Region	Submarine Cable %BW utilization	LIT Submarine Cable Capacity
Asia	69%	
Middle East	73%	
Oceania	56%	
Europe	52%	
US & Europe	51%	
Latin America & Caribbean	77%	
Africa	69%	
Transatlantic		15%

Cable	
US Latin America	37%
Transpacific	21%
Europe Asia	91%
Intra-Asia	31%

Submarine Cable Boom: In 2001 US $13.5 billion was spent. Due to Dot.com bust, only $2billion was spent during four lean years from 2004 to 2007. Even the during the lean periods, global demand never slaked growing at an average compound annual rate of 54% from 2002 to 2008. In Latin America and the Caribbean, traffic grew the most at more than 75%.

METALFE LAW: The value of a NETWORK grows as the square of the number of users.

Foot Note: 350 Disk Drive, introduced in IBM 305 Ramac Computer, consisted of 50 Number of 24-inch magnetically coated platters mounted on a single vertical spindle and rotating at high speed. In between the platters and looking rather like a giant animated hair-comb, was an assembly of read-write heads that clacked in and out, reading and writing data from and to the disks and passing the information to and from the machine's processor.

The drive was the size of two large refrigerators and was leased to customers at an annual rental of $35,000/-, which according to my calculations would be $250,000 in today’s currency.

But corporate customers thought it a bargain because it meant that their Main Frame Computers could become much more versatile and faster. A digital computer works by taking data from a permanent storage medium, carrying out operations on that data, and then writing back into storage.

Slowest part of this process was getting data transferred to and fro from the processor to the storage. Hard Disks offered a way of easing the bottleneck. The result was faster computation and faster data processing.

IBM’s colossal spinning plate rack held a total of 4.4MB of Data.

Today the hard disk in the iPod is just 1.8” in diameter and yet it can store 60GB of Data, which is almost 14,000 times the capacity of the 350 Disk File. The drive in laptop is 2.5” in diameter and has a capacity of 120GB.

This cheap and boundless mass storage has enabled the existence of companies such as Google, Amazon and eBay and services such as Apples iTunes, Wikipedia and Internet Archive.

The story of computing has been told almost entirely in terms of advances in processors and networks. Whereas in fact the success and proliferation of computer usage is as much due to the vast, fast, cheap mass storage provided by Hard Disks.

Development in Memory Technology: The convergence of consumer,computer and communication electronic system is leading to exponential growth in need of memories. Memories are required for code storage, computing input-output storage and data storage. In the past, we could associate a memory technology to a specific market segment : RAM to computer; NOR Flash memory to mobile communication; NAND Flash memory to consumer DSC; Today such distinctions are being blurred.A common format is emerging. The new electronic systems stack different non-volatile memory and xRAM and use microprocessor for facilitating interfacing and managing the overall memory.

Manual	operator directs the call
Automatic Dialling: *Strowger-rotary switches(UAXs,SAXs,TXS) *Electromechanical crossbar(TXKs) *Electronic-control reed-	2 letter exchange-2 to 5 digit number; 3-3-4 system adopted(area code-exchange prefix-line number); The number system 3-3-4 first used in

relay switches(TXE2) *SPC non-digital(TXE4) *Digital(Systems X and Y-AXE10: may be further categorised as to whether they are ISDN -capable, I suppose;5ESS; DMS100;UXD5)	Wichita, Falls, Texas.
Global Telephone System	Service switching points, mobile switching points
SS7(Signaling System 7) links telephone networks all over the globe. Every Telecom carrier and manufacturer has to work and develop around SS7	Smart coordinator that helps callers find one another
VoIP(voice on internet protocol)	It is a telephone service on Data Network(no taxes hence cheap)
Futuristic Systems (21CN-21st Century Network)	System will know how and where to locate the destination; Convergence-lots of systems but few devices; preferred formats - Voice or Video, Real Time or dekayed, reading the voice mails and hearing the emails

Evolution of Telephone Exchanges.

The onward march of Third Information Revolution in the first decade of 21st Century.
This is the continuation of the Preface of Solid State Physics and Devices - The harbinger of the Third Wave of Civilization.

The onward march of the Third Information Revolution in the first decade of 21 st Century.

Top 11 Technologies of the decade (2000-2010)

[From IEEE Spectrum, Vol 48, Number 1 International, January 2011]

Technology 1_SMARTPHONES:

In 1945 ENIAC was built. It occupied a living room. It took next fifty years (1945-1995) to bring a Main Frame Computer to a suitcase size. It took another 10 years to evolve to Smartphone size with all the features of a personal computer.

Smartphones have the features of music players, remote controls, gaming consoles. Today in 2011 they are serving as hot spots. They keep us connected to information.

In 1973 the first cell phone was invented by Motorola’s Martin Cooper with the intention of having a portable phone. But this portable cell phone became the platform for innovation upstaging the PC. Smartphone is used less as a phone and more as internet browser and game console. The tremendous advancement in hardware, software and network infrastructure has made the evolution of cell phone to smart phone possible.

The hardware of the smart phone is the following:

1. High resolution displays with touch screens or QWERTY keyboards;
2. Tiny camera lenses on the outside;
3. GHz processors;
4. Radio antennas and
5. Image sensors on the inside..

Development of Operating System(OS) was as important in the evolution of

Smartphone. It supported third-party applications and in the process extended phone capabilities far beyond what any manufacturer could provide. Smart phones can offload data and processing to the cloud servers. Current smart phones have several options: Wi-Fi and 3G. It gets connected to the best available network.

For rural and impoverished areas they are self contained unit. Neither additional network infrastructure is required nor reliable power is required. Its own battery is sufficient.

In 2007 the sales of smart phones surpassed sales of laptops and by 2014 smart phones will be the dominant medium for internet browsing.

Today text messages have evolved to twitter updates and voice chats have evolved to video conferencing. Soon we will be gathering information from wireless sensors in our environment and as well as from internet, interpreting them and deriving valuable feed back.

Eight Milestones in evolution of Phone to Smartphone:

First Milestone: IBM Simon Personal Communicator in 1993- it was a touch screen phone, with a calendar, address book, calculator, email and fax capability. Bulky , hefty and price $899.

Second Milestone: Nokia 9000 Communicator in 1996- it was a sandwich of Personal Digital Assistant and Mobile Phone.

Third Milestone: Ericsson R380 in 2000 – This was the first real Smart Phone. It used Symbian OS. This was the only OS for the Smart Phones but lately it is losing ground to Apple’s iOS and Google’s Android.

Fourth Milestone: Sony Ericsson P800 in 2002 – This had added features such as MP3 player, camera and colour touch screen.

Fifth Milestone: RIM Blackberry 6210 in 2003 –

Research in Motion had two-way text paging. It had QWERTY keyboard, emailing facility, basic web browser and it was fully integrated phone(earlier version required a headset.

Sixth Milestone: PALM TREO 600 in 2003 – It had Palm OS and supported third party applications. It had integrated functions like it could maintain and use a contact list. It could check the calendar while using the phone.

Seventh Milestone: Apple iPhone in 2007 – It had multitouch, capacitive touch screen which allowed interactive use of the phone. It had Safari browser which made mobile Web browsing very convenient. Apple was able to get many concessions from Mobile Ecosystem which made it possible for Apple i3G to upgrade the data connection and it was able to launch the wildly successful application store.

(Mobile Ecosystem is a leading consulting and advisory firm focused on next generation mobile communications products and services. The firm is led by Mark Lowenstein, who is widely recognized as a preeminent industry executive and thought leader. Lowenstein most recently served as VP, Market Strategy at Verizon Wireless. During his fifteen-year consulting career, Lowenstein was Executive VP at the Yankee Group prior to founding Mobile Ecosystem.)

Eighth Milestone: HTC EVO 4G in 2010 – It used Google OS Android. It had 8-Megapixel camera that could shoot high definition (HD) video and generate HDMI output. A HDMI high speed cable can be used for connecting the phone to HDTV and replay the video image on HDTV. It had 4.3-inch touch screen. It can act as mobile hot-spot connected to Sprint's WiMax network, the fastest data network in USA at the time.

Technology 2_SOCIAL NETWORKING

With a few click of the mouse contacts can be made and meet like minded people. This is social networking. It started with SIXDEGREES in 1997 and today it has evolved to FACEBOOK.

In 1997, 28 year old New York businessman Andrew Weinreich launched SIXDEGREES at a party. Its users identified their friends, their friends' friends and so forth. By 1999 it had 3.5 million members but this network had no provision for exchange of photographs.

At this time point-and-shoot camera were ubiquitous and broad band was cheap and available. In March 2003, FRIENDSTER was launched. A user could set up a profile, with personal facts and a picture, and invite friends to join. Friendster server would generate list of his friends as well as friends' friends, mapping relationships out to four degrees of separation. But it was taking customers faster than it could manage. As a result people started moving to MySpace.

MySpace had a meteoric success. It had 72 million members in 70 countries by 2007. The reign of MySpace was short lived. In 2008, it had 67% of American Social Networking market. By September 2009 it came down to 30%. It could not develop tools that would help users organize the vast amount of shared information. This made the site too congested and people left for something simpler.

In 2004, Mark Zuckerberg started FACEBOOK from Harvard University and expanded University by University. It had sleek, easy-to-use interface. It could keep updating its users' home page with information about their friends by continuous searching of the data base. Today Facebook contends with 30 billion shared updates a month. Today Facebook has 540 million users who spend 700 billion minutes on the site every month.

With the advent of Smartphones and Cloud computing, we are heading into the third phase of the Web. In first phase we had webs between humans and machines. Then we had web around humans and people they trust. And now we are having the web spun around our physical world and virtual world.

These digital tools such as Facebook and Twitter have been used by the people of Tunisia and Egypt to organize peaceful democratic protests which have led to regime change in a bloodless manner.

Technology 3_VOICE OVER IP(VoIP)

Timeline of Internet Telephony.

Year	Evolution of Internet
1970	ARPANET experiments with digital signal processing.
1974	The IEEE publishes a paper titled “A Protocol for Packet Network Interconnections”.
1995	VocalTec launches iPhone, the first commercial VoIP application.
1999	Mark Spencer creates Asterisk. Asterisk is an open source private branch exchange (PBX) originally created by Mark Spencer of Digium.

	A PBX, is a type of phone switch, that allows multiple attached telephones to make calls to one another, and to connect to other telephone services including the public switched telephone network (**PSTN**). Not totally unlike routers that connect multiple computers to a network.
2002	Vonage is launched.**Vonage** is a publicly-held commercial voice over IP (VoIP) network and SIP company that provides telephone service via a broadband connection. The company's name is a play on their motto "Voice-Over Net AGE".[
2003	Skype is launched. **Skype** (pronounced /ˈskaɪp/) is a software application that allows users to make voice calls over the Internet. Calls to other users within the Skype

	service are free, while calls to both traditional landline telephones and mobile phone can be made for a fee using a debit-based user account system. Skype has also become popular for its additional features which include instant messaging, file transfer, and video conferencing. The network is operated by Skype Limited, which has its headquarters in Luxembourg and is minority owned by eBay. Most of the development team of Skype is situated in Tallinn, Estonia for cost reasons and outsourced by the Luxembourg parent company[1]
2004	Microsoft Xbox Live is launched. REDMOND, Wash., Jan. 20, 2005 — Coming off one of the hottest holiday seasons for video game software sales in history, Microsoft Corp. today announced

	that its Xbox *Live (TM)* online gaming service has set a new standard with more than 1.4 million members, exceeding expectations and cementing its leadership position in the digital entertainment landscape. This additional news comes on the heels of 14 consecutive months of increased market share for Xbox® , the only console to see an increase of market share in 2004.
2007	VoIP over Wi-Fi applications appear for smartphones.
2009	VoIP over 3G applications appear.
2010	First LTE deployments take place. LTE or Long Term Evolution, is a a 4th generation (4G) mobile broadband standard and is aimed to be the successor to the 3G technologies GSM/UMTS. It is currently in development and is

considered the competitor to WiMax. Carriers will include Verizon, T-Mobile, Vodafone, AT&T, and many more worldwide. This technology will provide broadband services wirelessly (like EVDO), and will transmit signals via radio platform. You will need an LTE modem to access the network, which can be in USB format, ExpressCard, PCMCIA, or embedded in a laptop. It will also likely be featured as the internet connection on PDAs and phones. This network, which is promising peak download rates of up 100Mbps, will provide an alternative to DSL, cable, satellite, and dial-up internet. It will free people from the burden of having to find a WiFi hotspot when they are on the road – as long as you have an LTE modem, you can connect to the internet anywhere in the service provider's coverage area.

In bygone ages , you could never outgrow the handset of your youth. Telephones were tethered to the wall, trunk calls would remain a costly affair. But with the advent of Internet Telephony, cell phones and competition between service providers, all that has changed beyond recognition as the above time line shows. Still switch over to VoIP will not be happen overnight. 4G will coexist with 3G and 2G for much of the next decade. "But as communications adopt an all-IP architecture, it will get easier to overlay voice on top of other technologies. Digitized voice is finding its way into dozens of applications, including social networking, online gaming, videoconferencing, even advertising."

Technology 4_LED LIGHTING

LED was first introduced in 1970 as red LED and its application was confined to as power indicators. But as R&D progressed in the field, its efficiency , brightness and affordability improved. Today it is 20 times as bright and 90% cheaper per Watt. This improvement is known as Haitz's law. Later on yellow and blue LEDs were introduced and here also improvement took place according to Haitz's law.

Table 1. Progress in LEDs since 1970

	Status in 1999	Present status in 2010
Operating Current	10mA	1 A
Power consumption for 10 lumens of light	1W	100 lm/W
Physical size	300μm × 300μm	1 mm^2
Price	high	high
Life Time		25 times as compared to incandescent lamp and 150% longer life as compared to that of CFL.
Toxic content such as mercury		LED has no toxic content.

As shown in Figure 1, the first generation white LED was as shown but it had serious problems of light emission. Only a small portion of radiation was coming out of the front portion. To overcome this problem processing was changed.and second generation LED was commercialized. White LEDs were a commercial success. A 7W LED could deliver as much illumination as 40W incandescent lamp. A good quality LED would cost $55 but its investment would be recovered in 5 years. The manufacturing cost is going down as the substrate size is increasing and as yield is improving. The substrate size has increased from 2” to 3” to 4” and now 6”. The efficiency of energy conversion is being improved

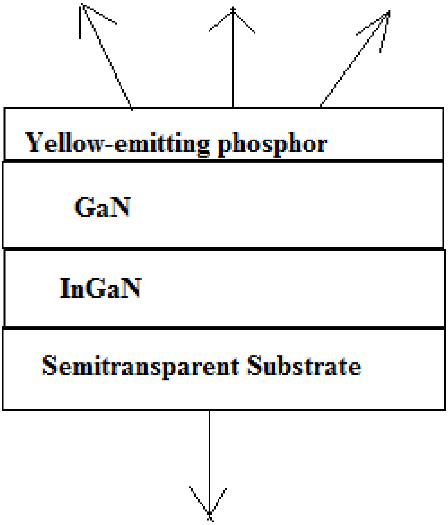


Figure 1. First generation White LEDs. GaN & InGaN emit blue light. In combination with yellow-emitting phosphor the structure becomes white LED but it is not an optimum configration. A large part of the emitted light gets trapped within the chip and a good fraction of light gets radiated the back of the LED through the substrate.

also. Right now 100 lumens of illumination per Watt has been achieved. By 2015, 200 lumens per Watt is expected to be achieved. This technology is destined to replace all the existing illumination technology.

Technolgy 5_MULTICORE CPUs

By scaling the transistors, computation speed could be improved . But this led to the problem of heating hence instruction level parallelism was introduced. This was known as “superscalar architecture”. Here quadrupling the physical resources led to linear increase in computational performance. At this point dual core processor was introduced. This achieved parallel processing without delay inducing complexity. For thermal power dissipation management, most of the processor giants such as Intel, AMD and IBM went for multicore processors. In 2005 Intel marketed first dual-core component, Pentium D. Here there were two single-core processors tied together by front-side bus in the same package.

IBM released Power4 in 2001. It had two cores on a single chip. In 2005, SunMicrosystems commercialized eight-core UltraSpare T1(also known as Niagra).

Today 50 percent annual improvement is achieved in effective processing power by parallelism and not by speedier transistors. Memory Bandwidth will be the constraint on multicore technology. No more than 16 cores can be realized on a single chip.

In a more advanced approach, CPU is being combined with Graphic Processing Units(GPU) into Accelerated Processing Unit(APU). This has been implemented by AMD.

Future trend is to combine general purpose cores with specialized cores such as encryption core, decryption core, videoencoding core or anything with well-defined standard.

Technology 6_CLOUDCOMPUTING

In 1970, Advance Research Program Agency(ARPA) established ARPANET, the first wide area network(WAN), connecting Stanford Research Institute, University of Utah, University of California Los Angeles and University of California Santa Barbara. This ARPANET was used for Defence Research Programs. In three years it grew to cover whole of USA.

In 1971, electronic mail or e-mail is invented by Ray Tomilinson. He combined the existing mail program ‘SENDMSG’ that worked only within an organization with a file transfer program called ‘CPYNET’ to create an e-mail utility between 15 computers within the organization and use the same on the existing ARPANET.

In 1974,TCP/IP (Transmission Control Protocol/ Internet Protocol) model and protocols were developed and standardized by Vint Cerf and Kahn.

In March,1989, Tim Berners-Lee of CERN, Geneva, handed a proposal ‘Information Management: a proposal’. The proposal came up with hypertext language , the basis of “http” in website address. By October 1990, this team developed the first web browser. The World Wide Web Technology was made available for wider use on INTERNET from 1991 onward without any royalties. “ Internet is a vast network of networks, interconnected in many different ways yet they speak the same language. Web is one – albeit the most influential and well known – of many different applications which run over the INTERNET.

In 1991, ARPANET was confined to exclusive clientele (Department of Defence Research). To make the NET available to the University Community at large, National Science Foundation connected its own NSFNET to ARPANET. This led to exponential growth in the use of NET. In mid-90s the collection of networks was looked at as INTERNET.

With the development of HTML(Hyper Text Markup Language) for easy file transfer and the subsequent building up of World Wide Web with graphical browsers and clickable hyperlinks, internet became of use for the whole population.

By 2000, blogs, tweets and social networking were being built providing information, entertainment and 24 hours connectivity. It is this connectivity along with Wikileaks Cable exposure which has led to Egyptian and Tunisian revolution.

The breaking down of the barrier between hardware and software has made possible the internet miracle. Now most of our computational data , our files and pictures can be saved on the servers provided by our service providers. We need not be worried about our computers crashing and losing all our data. This is known as Cloud Computing. The powerful convergence of broadband access, the profusion of mobile devices enabling near-constant Internet connectivity and the hundreds of innovations for building and running data centres has made possible Cloud Computing. Just as we had time sharing of Main-frame computers in 1970s, in exactly the same fashion through internet we can carry out intensive computational tasks on the servers of some service provider but this computational results will be accessible to all.

“The internet may sometimes seem like a massive astral projection,........... Not just Google and Amazon, but Microsoft, Rackspace, SGI and many other companies have used striking economies of scale to acquire cheap commodity hardware and build sprawling, multihectare server farms.

“Twitter, Facebook, Google – their performance needs rise astronomically everyday”...... To keep up , companies buy servers by the shipping container with all the power and networking connections preconfigured. An IT manager plugs the container into a power supply, a network connection, and a cold water tap, and presto, 1400-odd more servers come on-line.”[Sandra Upson, ibid]

In 2008, Google launched App Engine (application engine). Any user can upload his set of Python or Java Codes on Google App Engine site. These are modified and run on any number of desired machines. Any Web Application which becomes highly popular will be sensed and quantum of computing time is increased for your Web Application.

Cloud gives you the flexibility which is not found in any corporate offices. 10 servers can be run for 1000 hours at the same price as 1000 servers can be run for 10 hours.

Just as we have Electric Public Utility Service, in exactly the same manner we have Computational Public Utility Service. User will pay for what they use and don’t have to install their computational machines.

“The now-routine tasks of cloning virtual machines and distributing them through multiple data centers make for easy backups. And at a few cents per CPU-hour, cloud computing can be cheap as dirt”.

Data stored in the clouds must be safe and indestructible. For this cloud to cloud smooth transfer of data should be assured.

In the second decade of 21st century, we will see how all the computer functions are carried out by the internet as one entity .

Technology 7_DRONE AIRCRAFT

In May 1898, at the Electrical Exposition in New York City’s Madison Square Garden the concept of a remotely piloted vehicle was demonstrated. Nikola Tesla demonstrated the concept of “telautomaton” - small boat operated remotely by radio. This could be used as a torpedo against the enemy ship. In that year itself, Tesla sent a paper titled “Remotely controlled aircraft as an aerial torpedo” to The Electrical Engineer magazine. This paper was rejected as too fanciful a paper.

As early as 1917, the US Navy pursued the idea of a pilotless aircraft for use against German U-boats. During the Second World War, Nazi forces covered the British skies with thousands of pulse-jet-powered flying bombs. In 1982, Israel used drones against Syrian forces in Lebanon.

In the first decade of 21st century, these pilotless vehicles have matured into fully controllable and reusable combat aircraft. Their miniaturized forms act as easily deployed eyes in the skies for the ground forces.

Over the years, unmanned drones have evolved into unmanned aerial vehicles(UAV). These are also called unmanned aircraft systems (UAS). These UAV can function in a variety of ways as shown in the Table 2.

Table 2. Varied functions of UAV.

UAV	Functions
do	Reconnaissance or sustained surveillance
do	Illumination of targets by laser for destruction by guided weapons launched from other platforms.
do	Missile carrying UAV which will deliberately crash into the target and destroy it.
do	Portable UAV which can be manually launched.

Raven manufactured by AeroVironment, Monrovia, California	Radio-controlled planes like model crafts but is packed with sophisticated electronic gears.

Table 3. Improvements leading to an explosion in applications of UAVs.

UAV	Improvements made for making UAV effective
Raven.	Nickel-cadmium (gave power for 30 minutes) was replaced by lithium ion-powered batteries which lasted for 90 minutes in UAV.
	MEMS (microelectromechanical systems) was the enabling technology which allowed tiny gyroscopes, accelerometers and air-speed sensors to be added to the smallest of these vehicles.
	Compact and reliable GPS

	(global positioning systems) receivers.
	Advances in computers and radio links.
	Weapons could be used by 2001 in reusable UAV
1990s Predators.	Used only for surveillance.
Predators outfitted with laser designators.	These could automatically guide missiles or artillery shells launched from ground platforms.
Predators as air-to surface missile launchers.	Used in October 2001 in the hunt for al-Qaeda members in Afghanistan
Predator as logistic support	Used in destroying a machine gun bunker that had pinned down U.S. Army rangers in Afghanistan.
Predator as air-to-air missile launcher	Used in 2003 against Iraqi fighter.
‘Shadow’ manufactured by A.A.I. Corp., the Hunt Valley, Madison.	Special radio-equipments added so that by automated landings were possible even in blinding sandstorms.

The possible civilian applications.

1. UPS and Fedex flight;

2. Pipeline surveys;
3. Logging;
4. Law enforcement;

Advances in communication and automation technology will help in the further proliferation in UAV applications.

Technology 8_PLANETARY ROVERS.

Before we describe the remarkable advances in Planetary Rovers, it will be appropriate to describe the ushering in of Space Age by the launching of Sputnik.

Table 4. The unfolding of Space Age - A BRIEF HISTORY OF APPOLLO AND LUNA MISSION*

4th Oct. 1957	Sputnik (I) launched by erstwhile U.S.S.R. It carried a Radio Transmitter. This was an aluminium sphere 58 cms in diameter and 83.6 kg in weight. It measured atmospheric density and temperature throughout

	its elliptical orbit which ranged from 227km at its perigee to 941 km at its apogee. It also collected data on electrons concentration in ionsphere. The satellite was launched by a modified R.7(55-6 Sapwood), the first Soviet Intercontinental Ballistic Missile(ICBM).
3rd Nov. 1957	Sputnik (II) launched. It carried a dog, laika. The space endurance of the dog, in-situ, was monitored by methods of Telemetry.
17th April 1961	Yuri Gagrin, the first Cosmonaut, in Space atop VOSTOK (1). Duration of stay in Space was 108 minutes.
May, 1961	President of U.S.A. makes his historic Statement :“I believe this Nation should commit itself to achieving the goal, before the decade is out, of landing Man on Moon and returning him safely to

	Earth.”

- Seigei Korolev was the major architect of Soviet Space Programme. He had designed Sputnik 1.

APOLLO MISSIONS.

Dec. 24, 1968	Apollo 8 carried out LUNAR ORBITAL MISSION. Moon’s surface was photographed and sent back by digital technique. The minimum altitude over Moon was 115 km.
May 21, 1969	Apollo 10 repeated Lunar Orbital Mission but achieved a minimum altitude of 17 km.
July 20, 196920 : 17 GMT	Apollo 11 carried out the historic mission of putting Man on Moon. Saturn V rocket launched Apollo Spacecraft on the course to Moon. The crew consisted of Mission

Commander, Neil Armstrong, Pilot of the Lunar Module, Eagle, Edwin E. Aldrin, and Command Module Pilot, Michael Collins. Apollo 11 had landed in Sea of Tranquility, a low land. 22 kg mass of Moon rocks and samples brought back. Eagle while trying to land traveled westward overshooting the target by 6 km and 2 km off the track due South. Eagle was supposed to land on smooth patch in the Sea of Tranquility. Instead it landed in a most inhospitable terrain full of boulders. This actual landing site was outside the region that had been chosen for and catered for in ground simulation of landing. This mishap occurred because of sketchy knowledge of Moon’s gravitational field. Mass concentrations or MASCONS cause perilun

	wiggle which have to be accounted for in trajectory calculation. These MASCONS are either remnants of the asteroids that hit the Moon or dense material from the interior. If these wiggles are not accounted for then trajectory calculation will go haywire as it did in the case of Eagle's landing.
Nov. 18, 1969	Apollo 12 landed in Oceanus Procellarum (the ocean of storms). Sample of Mare Basin was brought back. The participants were Pete Conrad and Allen Bean. The landing site was 200 meters from the landing site of Surveyor 3 probe. Apollo 12 landed exactly on target.
	Apollo 13 mission got aborted and had to come back to Earth.
Feb. 2, 1971	Apollo 14 landed on Fra Mauro (ejecta from Imburin Basin). Samples

	of ejecta from Imburin Basin were brought back. Situ Rossa piloted the Command Module and Alan Shepard and Ed Mitchell explored Fra Mauro highlands.
July 30, 1971	Apollo 15, landed at the edge of Mare basin. In this Mission, Lunar Buggy, a roving vehicle, was used. Samples of material from Apennine Mountains, the rim of Imburin Basin, was brought back. Apennine mountains are Moon’s primordial crust. Samples of Mare rocks and dust were brought back. Imburin Basin is the great canyon on Moon. David Scott and Jim Irwin spent three days on Moon. In these three days thrice they drove their Lunar Buggy and explored Moon’s surface for seven hours in each drive/
April 20, 1972	Apollo 16 landed on Lunar uplands near crater

	Descrates. Samples of highland brought back.
Dec. 11, 1972	Apollo 17, landed on Taurus mountains. Samples of a region suspected of recent volcanism was brought back. Rim Evans piloted the Command Module. Gene Cornan and geologist Jack Schmitt did 3 days of prospecting in Taurus Littrow Valley.
	In all 24 men participated in Apollo Mission and 12 men, amongst these, set foot on Moon.

LUNA MISSIONS

Feb. 1966	Luna 9 carried out soft landing Mission. It sent back detailed pictures of hard and rocky landing site.
March 1966	Luna 10 carried out Lunar

	Orbital Mission. It sent back data on radiation revels and micro meteorite impacts.
1970	Luna 17 Mission carried out. *Lunokhod 1, a roving vehicle,* was placed on Moon’s surface. Lunokhod 1 had an active life of 10 months and traveled 10.5 km. Enroute it did sample testing and sent back data by telemetry.
1973	Luna 21 mission carried out. *Lunokhod 2* was placed on Moon’s surface. This vehicle had a life of 4 months and traversed 37 km. Enroute this also did sample testing. By television/radio link remote control of the roving vehicle and the mechanical arm was carried out and through telemetry techniques communication maintained.

As we see from the above Table 4, in the early part of space programmes Lunokhod 1 and Lunokhod 2 placed by Luna17 in 1970 and by Luna 21 in 1973 were the first Planetary Rovers to be used.

As interest moved away from our Moon to more distant parts of the Solar System, flyby probes, orbiters and landers came to dominate the scene. In mid-1970s, NASA's Viking Program came into being.

The Viking program consisted of a pair of space probes sent to Mars—Viking 1 and Viking 2. Each vehicle was composed of two main parts, an orbiter designed to photograph the surface of Mars from orbit, and a lander designed to study the planet from the surface. The orbiters also served as communication relays for the landers once they touched down. Viking 1 was launched on August 20, 1975, and the second craft, Viking 2, was launched on September 9, 1975, both riding atop Titan III-E rockets with Centaur upper stages. By discovering many geological forms that are typically formed from large amounts of water, the Viking program caused a revolution in scientific ideas about water on Mars.

The primary objectives of the Viking orbiters were to transport the landers to Mars, perform reconnaissance to locate and certify landing sites, act as a communications relays for the landers, and to perform their own scientific investigations. The orbiter, based on the earlier Mariner 9 spacecraft, was an octagon approximately 2.5 m across. The total launch mass was 2,328 kilograms (5,130 lb), of which 1,445 kilograms (3,190 lb) were propellant and attitude control gas.

The landers were looking for signs of life which they didnot find mainly because they were stay put in one place.

Rovers made a comeback with NASA’s Pathfinder program .The *Mars Pathfinder* (MESUR Pathfinder, [34]) later renamed the *Carl Sagan Memorial Station*, was launched on December 4, 1996, just a month after the *Mars Global Surveyor* was launched. Onboard the lander was a small rover called *Sojourner* [Figure 2] that would execute many experiments on the Martian surface. It was the second project from NASA's Discovery Program, which promotes the use of low-cost spacecraft and frequent launches under the motto "cheaper, faster and better" promoted by the then administrator, Daniel Goldin. The mission was directed by the Jet Propulsion Laboratory, a division of the California Institute of Technology, responsible for NASA's Mars Exploration Program. This mission, besides being the first of a series of missions to Mars that included rovers (robotic exploration vehicles), was the most important since the *Vikings* landed on the red planet in 1976, and also was the first successful mission to send a rover to a planet. In addition to scientific objectives, the Mars Pathfinder mission was also a "proof-of-concept" for various technologies, such as airbag-mediated touchdown and automated obstacle avoidance, both later exploited by the Mars Exploration Rovers. The Mars Pathfinder was also remarkable for its extremely low price relative to other unmanned space missions to Mars.

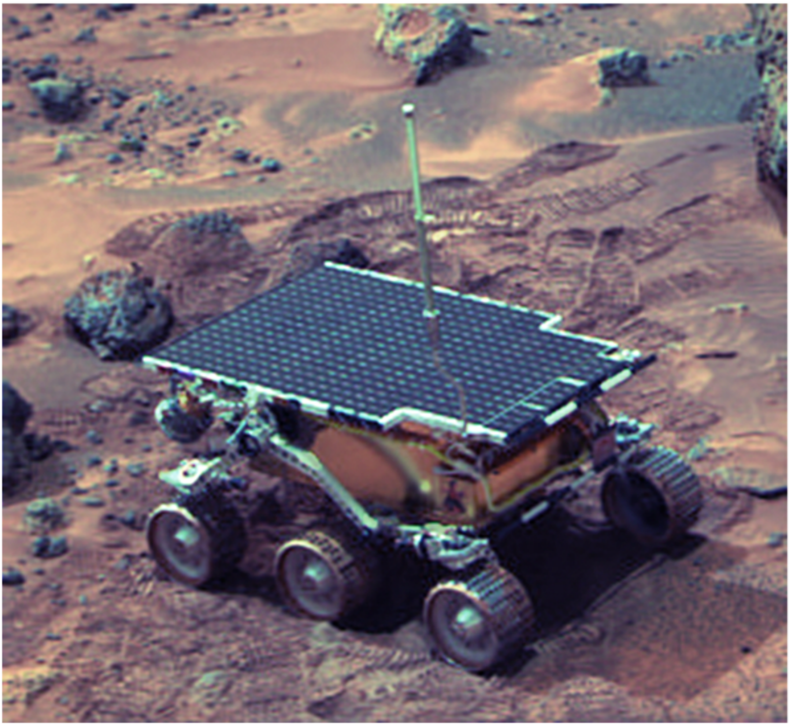


Figure 2. The Soujourner rover on Mars

In 1999 NASA lost two space crafts Mars Climate Orbiter and the Mars

Polar Lander .

At this time after careful review, NASA used Pathfinder’s successful

landing approach to place not a lander but a full size rover on Mars. This saw the start of

Mars Exploration Rover Mission.

NASA's *Mars Exploration Rover Mission* (MER), is an ongoing robotic space mission involving two rovers exploring the planet Mars. The mission is managed for NASA by the Jet Propulsion Laboratory, which designed, built and is operating the rovers.

The mission began in 2003 with the sending of the two rovers — MER-A *Spirit* and MER-B *Opportunity* — to explore the Martian surface and geology. The mission's scientific objective is to search for and characterize a wide range of rocks and soils that hold clues to past water activity on Mars. The mission is part of NASA's Mars Exploration Program which includes three previous successful landers: the two Viking program landers in 1976 and Mars Pathfinder probe in 1997.[35]

The total cost of building, launching, landing and operating the rovers on the surface for the initial 90-Martian-day (sol) primary mission was US$820 million.[36] Since the rovers have continued to function far beyond their initial 90 sol primary mission (to date both rovers have been functioning on Mars's surface for nearly seven years), they have each received multiple mission extensions.

As of now, 1st March 2011, Spirit is stuck in sand and probably has not survived Martian Winter but Opportunity is still roving around.

In recognition of the vast amount of scientific information amassed by both rovers, two asteroids have been named in their honor: 37452 Spirit and 39382 Opportunity.

Over the years it had become clear that unmanned space missions are the only viable financial propositions achieving comparable results as the manned missions at considerably lower risk. By the launch of ‘Spirit’ and ‘Opportunity’, the twin Mars Rovers launched in Mars Exploration Rover Mission in 2003, it has been demonstrated that

1. these are very effective exploration tools;
2. These have fostered scientific and technological innovations that should find applications on Earth;
3. These should find applications in areas such as autonomous robotics, remote sensing and materials engineering.

Today NASA(US Space agency), ESA(European Space Agency), Russian Space Agency, Chinese Space Agency and Indian Space Agency have Rover Technology as an essential part of their respective Space Programmes.

Jet Propulsion Lab (JPL) is NASA's lead facility for the exploration of the Solar System. Hardware built in this Lab has been flown to our Moon, Venus, Mars, Mercury, Jupiter, Saturn, Uranus and Neptune.

JPL has launched a new Rover Program called Curiosity Rover under the same Mars Exploration Rover Mission. It is one order of magnitude harder to design and build. It will have a massive 2.3-m-long arm. With this arm it will use a percussion drill to extract samples from within the Martian Rocks. Wheels half a meter in diameter will enable it to traverse difficult terrain, off-road style. With its captive power plant of plutonium-238 thermoelectric generator, it will be able to operate in the winter and at latitudes farther from the equator. The goal is to find micro-organisms in places where previous rovers did not go and to better understand how the availability of water, energy and elements like carbon evolved on the surface of Mars.

Curiosity will carry 10 instruments from 5 countries. The most important instrument is SAM (sample analysis at Mars). It has a mass spectrometer, a tunable laser spectrometer and a gas chromatograph to characterize the molecular structure and the isotopic composition of the soil sample it is examining and to test for organic carbon.

The science hardware will be altogether 80 kg as compared to 5 kg carried by Spirit and Opportunity each.

Curiosity will touch down on Mars surface in a most dramatic manner through a sky crane:

1. Curiosity will travel aboard a capsule known as aeroshell;
2. After entering the Martian atmosphere, a parachute will unfurl;
3. While plunging at supersonic speeds, five retro rockets will be fired to decelerate the aeroshell to a gentle descent;
4. At an altitude of 20 meters, rover will detach from the aeroshell and lower itself on cables;
5. When rover touches the ground, explosive bolt cutters will release the cables, allowing aeroshell to fly away and crash at a safe distance.

The launch of Curiosity is set for December 2011.

The ultimate objective behind this Rover Exploration Programme Is:

- To improve life here, to extend life to there, to find life beyond.[57] — Mission Statement
- NASA's mission is to pioneer the future in space exploration, scientific discovery and aeronautics research.[58] — Mission
- To understand and protect our home planet, to

> explore the Universe and search for life, and to inspire the next generation of explorers... as only NASA can.[57] — NASA Vision

Technolgy 9_FLEXIBLE AC TRANSMISSIONS

“Power Systems must juggle supply and demand while guaranteeing glitch-free alternating current 24/7…….”.

“[But] Over the past the past decade, the confluence of innovations, regulatory change, and sheer watt-squeezing necessity has hatched a marvellous advance, one that has begun to realize the long-standing dream of pushing current where it would not ordinarily go”.

Over the past decade Flexible AC Transmission System or FACTS has come of age. Its objective is to maximize the throughput with minimum of losses. This objective is achieved by Smart Grid. This helps reconfigure power flows in real time according to the real time demand. This is what helps attain the objective of maximum throughput and minimum loss. This also helps incorporate wind, solar and other intrinsically intermittent sources of energy into the National Power Grid smoothly.

What exactly is meant by Smart Grid:

1. Balancing active and reactive power and thereby improve power factor. In our good old

days we had two methods of doing this - first by putting bank of shunt capacitances into a circuit or by tuning the generators in conventional power stations to produce current waveforms that lead the voltage waveform.

2. Our loads are generally of inductive kinds: transformers, motors and generators;
3. Too much inductive load and hence inductive reactive power causes voltage to “sag”.
4. Too much of “sag” will damage the electrical equipment.
5. This damage can spread leading to cascading failures. This happened in 2003 blackout that toppled grids from Ottawa to Baltimore.
6. Smart grids match the supply and demand in a local region at an hourly basis. The supply can exceed the rating of the transmission power lines by almost 50% .
7. A modern first generation FACTS controller in 1970 had an array of solid-state switches often coupled with capacitors. Power is taken from the National Grid to charge the bank of capacitances and then switches fire to generate the requisite phase lead which will be optimally suited for compensating the phase lag in the Local Grid. This assessment is done on a dynamic basis and thereby Power Factor is kept at unity.
8. In 1990, high-power semiconductor switches were introduced that could switch at frequencies higher than the standard 50- or 60-

hertz AC cycle. With insulated-gate bipolar transistors, FACTS controller could regulate the voltage and remove a variety of glitches. One such controller is ‘static synchoronous compensator’ or statcom. The wind turbines give a very unruly output. This kind of erratically varying output both in amplitude as well as in frequency could easily destabilize the congested Grids. The gusty, noisy power signal from a wind farm was seen as a threat to grid stability. Statcom helped stabilize the output of these wind farms. There are 100 3-MW offshore wind turbines completed this September 2010. This power is destined to satisfy the needs of 200,000 homes in UK. This 100 × 3-MW is stabilized by statcom and then fed to the National Grid for its destination.

9. In 1998, third generation FACTS controller was introduced. This stabilized the line and reduced the apparent impedance so that a larger quantum of power could be fed down. Brazil commissioned a pair of 1000-km, 500-kV line linking Northern Grid to Southern Grid. Northern Grid was supplied by hundreds of Hydro-power plants whose power had to be stabilized and power fact had to be improved. This was done by third generation FACTS. Southern Grid supplied power to the coastal population centers. Prior to this the Grid Lines were limited to few hundred miles due to high impedance. Now with the use of FACTS the

span could be improved to 1000-km. Prior to this lower resistance High-Voltage DC line was used as long span line for providing power to the coastal areas. This HVDC line could not be tapped to serve the

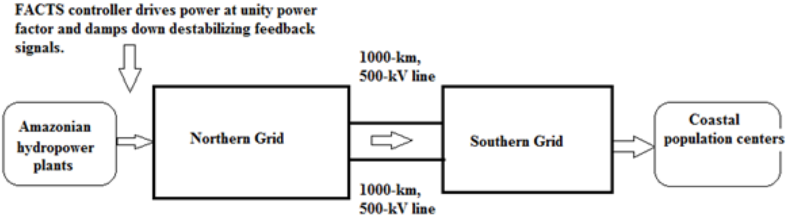

FACTS controller drives power at unity power factor and damps down destabilizing feedback signals.
1000-km, 500-kV line
Amazonian hydropower plants
Northern Grid
Southern Grid
Coastal population centers
1000-km, 500-kV line

Figure 3. In 1998 Brazil commissioned a pair of 1000-km, 500-kV lines to link its Northern Grid, replete with Amazonian hydropower, to Southern Grid serving its coastal population centers.

the interior communities. But the FACTS controlled line could be used to give electricity to hinterland of Brazil. Soon all the HVDC lines carrying wind power will be supplanted by this third generation FACTS controlled lines.

In India and China, FACTS controller are being used for using the power

plants at full capacity thus reducing the number of lines required to supply power to a given destination.

In 2003, Electric Power Research Institute and New York demonstrated that using FACTS controller , power could be diverted around a road-block from one Grid to another. But these kinds of power flow require higher efficiency of FACTS controller.

Table 5. Comparative power consumption by Insulated Gate Bipolar Transistor(IGBT) Switches used in FACTS Controllers.

IGBT	Maximum allowable temperature	Efficiency	Active Cooling requirement
Si-IGBT	100°C to 110°C	92 to 94%	MW consumed in active cooling
SiC-IGBT	200°C	96 to 97%	kW consumed in active cooling.

From the above Table 5, it is obvious that FACTS power routing entails a larger power “tax” in case of Si-IGBT.

The latest generation of HVDC converters are close technological cousins of advanced FACTS controllers and are also close competitors for two applications namely:

1. HVDC are more suitable for connecting off-shore Wind Farms to the mainland Grid;
2. HVDC are also more suitable for interconnecting AC systems which are adjacent but not synchronized.

The future vision of Smart Grid is that it should be flexible and scalable. The key to these two features is Power Electronics. So “ Power Electronics is the key Technology for the grids of the future”.

Technology 10_DIGITAL POTOGRAPHY

In 1957, the first digital image scanner was created at the US National Bureau of Standards. It was known as Facsimile or FAX images.

The most basic fax mode transfers black and white colors only. The original page is scanned in a resolution of 1728 pixels/line and 1145 lines/page (for A4). The resulting raw data is compressed using a modified Huffman code optimized for written text, achieving average compression factors of around 20. Typically a page needs 10 s for transmission, instead of about 3 minutes for the same uncompressed raw data of 1728 × 1145 bits at a speed of 9600 bit/s. The compression method uses a Huffman codebook for run lengths of black and white runs in a single scanned line, and it can also use the fact that two adjacent scanlines are usually quite similar, saving bandwidth by encoding only the differences.

In 1969, Charge Coupled Devices were invented and started storing the image.

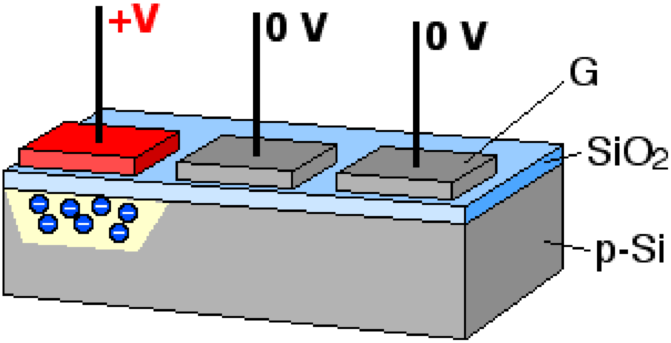


Figure 4. A three-phase charge coupled device.

Function of a three-phase charge-coupled device (CCD). The charge packets (electrons, blue) are collected in potential wells (yellow) created by applying positive voltage at the gate electrodes (G). Applying positive voltage to the gate electrode in the correct sequence transfers the charge packets.

An image is projected through a lens onto the capacitor array (the photoactive region), causing each capacitor to accumulate an electric charge proportional to the light intensity at that location. A one-dimensional array, used in line-scan cameras, captures a single slice of the image, while a two-dimensional array, used in video and still cameras, captures a two-dimensional picture corresponding to the scene projected onto the focal plane of the sensor. Once the array has been exposed to the image, a control circuit causes each capacitor to transfer its contents to its neighbor (operating as a shift register). The last capacitor in the array dumps its charge into a charge amplifier, which converts the charge into a voltage. By repeating this process, the controlling circuit converts the entire contents of the array in the semiconductor to a sequence of voltages. In a digital device, these voltages are then sampled, digitized, and usually stored in memory; in an analog device (such as an analog video camera), they are processed into a continuous analog signal (e.g. by feeding the output of the charge amplifier into a low-pass filter) which is then processed and fed out to other circuits for transmission, recording, or other processing.

Detailed physics of operation

The photoactive region of the CCD is, generally, an epitaxial layer of silicon. It has a doping of p+ (Boron) and is grown upon a substrate material, often p++. In buried channel devices, the type of design utilized in most modern CCDs, certain areas of the surface of the silicon are ion implanted with phosphorus, giving them an n-doped designation. This region defines the channel in which the photogenerated charge packets will travel. The gate oxide, i.e. the capacitordielectric, is grown on top of the epitaxial layer and substrate.

Later on in the process polysilicon gates are deposited by chemical vapor deposition, patterned with photolithography, and etched in such a way that the separately phased gates lie perpendicular to the channels. The channels are further defined by utilization of the LOCOS process to produce the channel stop region.

Channel stops are thermally grown oxides that serve to isolate the charge packets in one column from those in another. These channel stops are produced before the polysilicon gates are, as the LOCOS process utilizes a high temperature step that would destroy the gate material. The channels stops are parallel to, and exclusive of, the channel, or "charge carrying", regions.

Channel stops often have a p+ doped region underlying them, providing a further barrier to the electrons in the charge packets (this discussion of the physics of CCD devices assumes an electron transfer device, though hole transfer, is possible).

The clocking of the gates, alternately high and low, will forward and reverse bias to the diode that is provided by the buried channel (n-doped) and the epitaxial layer (p-doped). This will cause the CCD to deplete, near the p-n junction and will collect and move the charge packets beneath the gates—and within the channels—of the device.

CCD manufacturing and operation can be optimized for different uses. The above process describes a frame transfer CCD. While CCDs may be manufactured on a heavily doped p++ wafer it is also possible to manufacture a device inside p-wells that have been placed on an n-wafer. This second method, reportedly, reduces smear, dark current, and infrared and red response. This method of manufacture is used in the construction of interline transfer devices.

Another version of CCD is called a peristaltic CCD. In a peristaltic charge-coupled device, the charge packet transfer operation is analogous to the peristaltic contraction and dilation of the digestive system. The peristaltic CCD has an additional implant that keeps the charge away from the silicon/silicon dioxide interface and generates a large lateral electric field from one gate to the next. This provides an additional driving force to aid in transfer of the charge packets.

CCDs containing grids of pixels are used in digital cameras, optical scanners, and video cameras as light-sensing devices. They commonly respond to 70 percent of the incident light (meaning a quantum efficiency of about 70 percent) making them far more efficient than photographic film, which captures only about 2 percent of the incident light.

In December 1975, Steven Sasson at Eastman Kodak Corporation, Rochester, New York,

improvised a digital camera and did the first digital photography.

Table 6. Timeline of evolution of Digital Camera.

Year	Company	Type	Resolution	Comments
1975	Kodak	Digital, ADC-DAC is used	100- by 100-pixel CCD Chip	Digital data cassette recorder
1976				Pictures taken and shown on TV screen
1981	Sony	Analog	?	On floppy disk
1988	Kodak This was sent on shuttle mission Price $23000.00	Digital	Megapixel	

1991	KodakB&W Digital Camera System DCS100Price $20000 to $25000	Digital	?		The picture can be reviewed on electronic displays and sent on telephone lines using dial-up MODEM.Electron was separated from CCD Camera.
1992	Kodak DCS200Price $9000				All in one.
1994	Apple_Quick TakeUnder $1000Manufactured by Kodak			640 by 480 pixel, it could store 8 images	
1995	JPEG compression standard introduced				The huge data base could be compressed to

			reasonable size using JPEG which could be stored in then existing memory space and which could be transmitted within the Internet BW.
1995	Apple_Quick Take 150 with JPEG compression		
1995	Casio QV10		LCD
1997	Kodak_DC210	Megapixel consumer camera	
2001	Camera Phones in Japan		
2004	Camera Phones in USA		

Today the capability of camera is far outstripping the average consumer’s ability. The digital photo formats is rapidly changing. How do we store all the digital photos we are taking ?

Technology 11_CLASS-D AUDIO

We have seen that Class A power amplifiers power conversion efficiency is limited to 50% and Class B and Class AB is limited to 80%. For achieving higher efficiencies we had to switch to digital modulation format.

In digital modulation we have Pulse Amplitude Modulation(PAM), Pulse Width Modulation (PWM) and Pulse Position Modulation (PPM). Class D Amplifiers are PWM based amplifiers which have a much better conversion efficiencies (90% plus) and are suitable for portable and low power applications.

In digital modulation scheme we sample the analog signal at a frequency fs (sampling frequency) twice or greater than the maximum frequency (f_M) of the base band frequency as dictated by Nyquist Sampling Frequency Theorem in Pulse Width Modulation format as shown in Figure 5. The comparison of the audio signal with respect to the sawtooth waveform generates inverted PWM signal as shown in Figure 6. This signal is amplified by CMOS inverter. CMOS inverters work at a higher efficiency hence this has a much higher conversion efficiency. The PWM is passed through a Low Pass Filter to obtain the amplified analog signal. Since audio baseband has a highest frequency of 20kHz hence the sampling frequency is 250kHz to 1 MHz.

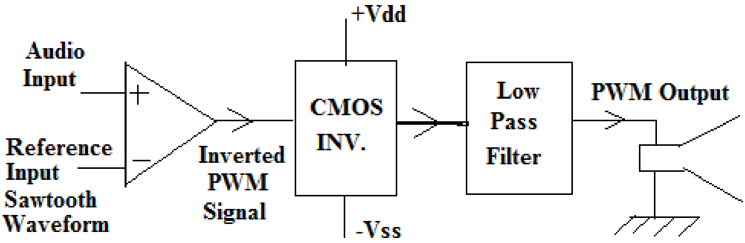


Figure 5. Schematics of Class D Amplifier.

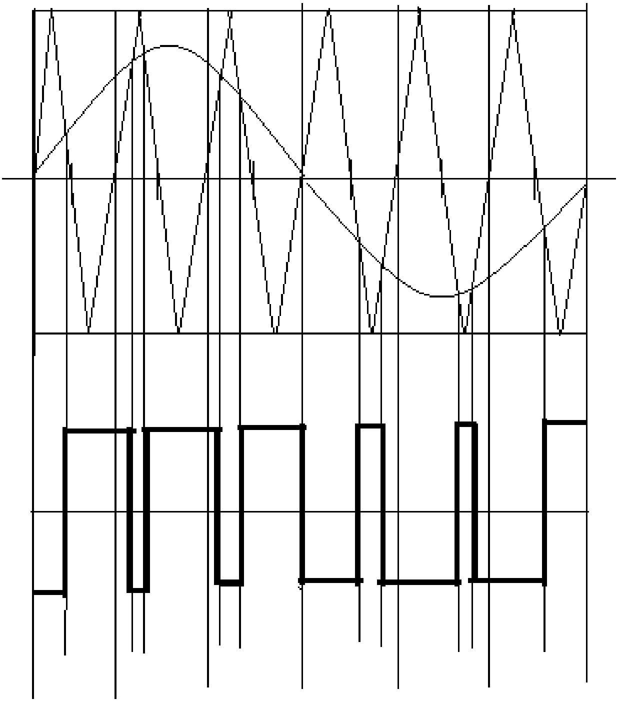


Figure 6. Generation of PWM signal by comparison of the audio signal with respect to sawtooth waveform

Class D has a superior performance and 90% plus power conversion efficiency hence today it dominates Audio Power Amplifier market.

Table 7. Time line of evolution of Class D Amplifiers.

Year	Company	Comments
1960s	British Sinclair Radionics	Performance poor
Early 1970s	Infinity Systems	Class D designed.
Mid 1980s	Peavy Electronics Corporation	Digital Energy Conversion Amplifiers- components not up to the mark hence high quality sound could not be produced- high speed sampling was not possible.
Late 1990s	Tact Audio, ICE power, Tripath Technology	MOSFETS available for high speed sampling.
1998	Tact Audio	At a price of

	renamed Lyngdorf	$9800 class D amplifier introduced called Tact Millennium.
1996	Tripath	First IC chip TA1101 of Class D amplifier introduced and used in Apple's Powr Mac G4 Cube(small form factor Macintosh Computer).
1998	Tripath	Second IC chip TA 2020- it could give an output of 20W per channel into 4 ohm speaker. Used in ministereos and early Flat screen TVs
2010		It dominates in micropower applications.
Future		It is going to be Class D world.

Moore Law continues in the second decade of 21st Century.
A paradigm shift takes place by the introduction of 22nm Tri-Gate Transistor which moves Transistor into third dimension and gives a new lease of life to Moore's Law in the second decade of 21st Century.

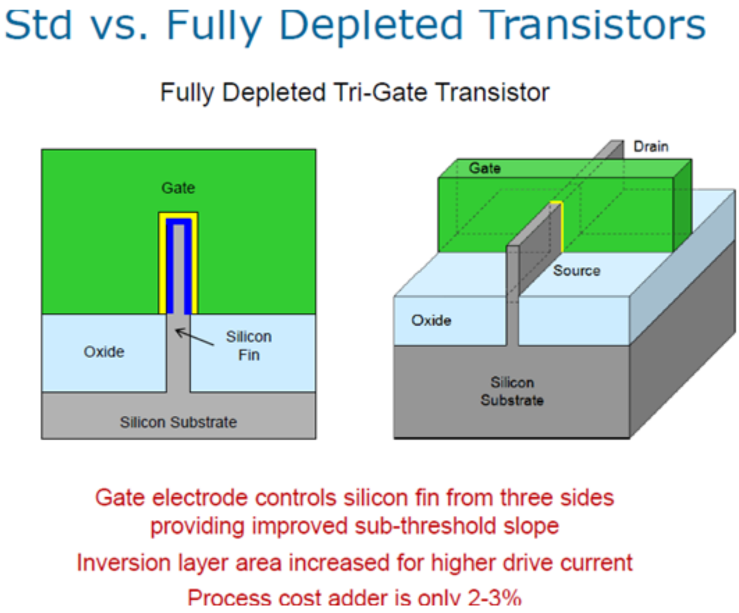
The Onward March of Information Revolution in the second decade of 21 st Century.

Intel’s 22nm Technology moves the transistor into 3 rd Dimension.

In the Journey of IC Technology from Micro to Nano Era we have seen the major developments which have taken place to allow the Moore’s Law to be followed by Industries in IC manufacturing. But as we scale down below 45nm Technology a paradigm shift was inevitable if Moore’s Law is to be obeyed in future.

On January 29, 2007, metal gate last and high-K technology in 45nm Technology was introduced. Using this technology Intel Core 2 Duo, Intel Core 2 Quad and multi-core processors were introduced.

On May 4, 2011, Intel introduced Tri-Gate Transistor in 22nm Technology. This was the much awaited paradigm shift which will permit Moore’s Law to be continued for another decade.In conventional 2D MOSFET, a conducting channel is formed in the silicon region below the gate electrode in ON condition. High-K dielectric and metal gate has enabled the sub-threshold current to minimize but it has not completely turned off the channel. Hence power wastage remains though reduced.



Std vs. Fully Depleted Transistors

Fully Depleted Tri-Gate Transistor

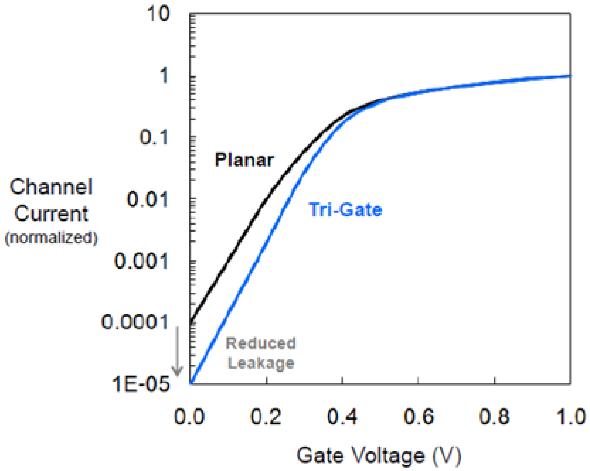
Gate electrode controls silicon fin from three sides providing improved sub-threshold slope

Inversion layer area increased for higher drive current

Process cost adder is only 2-3%

Figure 1. The cross-sectional view and the perspective view of Tri-gate Transistor.

In fully depleted tri-gate transistor, conducting channels are formed on three sides of a vertical fin structure as shown in Figure 1. The additional control (on the sides) enables as much transistor current to flow as possible when the transistor is in the “ON” state and it completely shuts down the current when it in “OFF” condition. This provides unprecedented combination of performance and energy efficiency. Hence it is suitable for a whole range of Computer Applications from servers to desktop and from lap-top to hand-held devices.



Transistor Operation

The “fully depleted” characteristics of Tri-Gate transistors provide a steeper sub-threshold slope that reduces leakage current

Figure 2. This I-V curve brings out the fully depleted character of Tri-Gate transistors.

This additional control speeds up the toggling rate also. The first microprocessor chip 4004, which had 2300 transistors and which used 10μm technology, toggled at 108kHz. In contrast Tri-Gate Transistor uses 22nm technology and switches 4000 times faster and it cuts down energy consumption by a factor of 1/5000 . The price per transistor has fallen down by a factor of 1/50,000 .

Intel’s 22nm Tri-Gate use third generation high-k/ metal-gate scheme, copper interconnects, strained Silicon and 193 nm Immersion Lithography. These terms have been discussed in IC journey from micro to nano era.

Because of heat management considerations 8086 architecture was being gradually replaced by ARM(Advanced RISC Machines) in CPU applications. But the use of 22nm Tri-Gate dramatically changed the energy efficiency scenario as a result 8086 will continue to be in race for CPU applications.

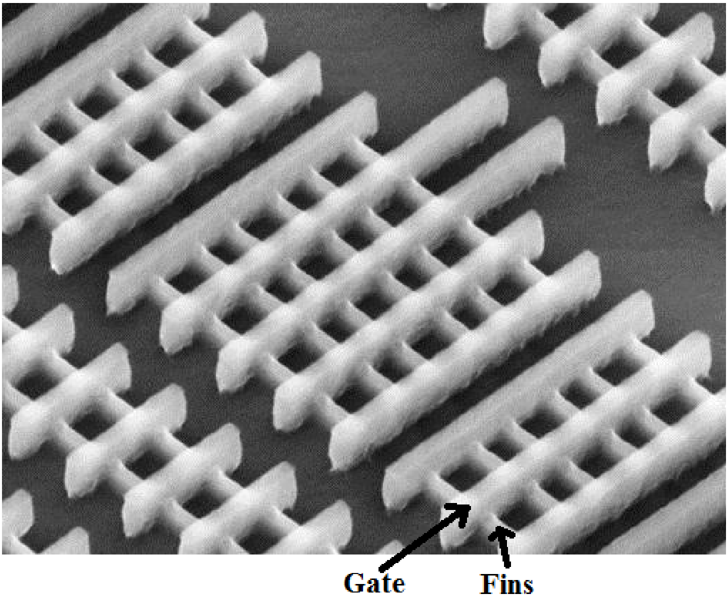
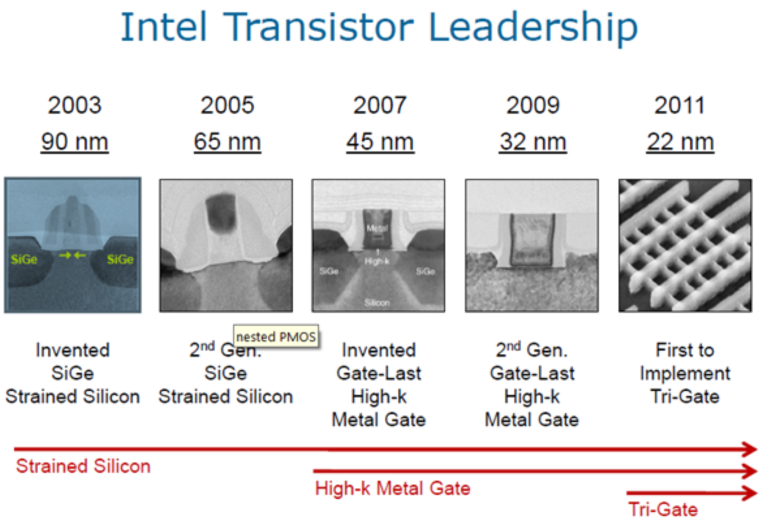


Figure 3. The photo-micrograph of Tri-Gate Transistors based IC.

Intel Technology Road Map.

Process Name	P1266	P1268	P1270	P1272	P1274
Lithography	45nm	32nm	22nm	14nm	10nm
First Production Year	2007	2009	2011	2013	2015



Intel Transistor Leadership

2003 90 nm	2005 65 nm	2007 45 nm	2009 32 nm	2011 22 nm
Invented SiGe Strained Silicon	2nd Gen. SiGe Strained Silicon	Invented Gate-Last High-k Metal Gate	2nd Gen. Gate-Last High-k Metal Gate	First to Implement Tri-Gate

SiGe
SiGe
nested PMOS
Metal
High-k
SiGe
SiGe
Silicon
Strained Silicon
High-k Metal Gate
Tri-Gate

Compared to 32nm 2-D transistors, Tri-Gate transistors will be 37% faster and power efficiency will be 50% better. That is for the same operating conditions, Tri-Gate will consume 50% less power. Because of better power efficiency, battery back-up will be prolonged and battery-life will be improved.These will be run at lower voltages. This will also add up to the back-up time as well as the battery life. This will be a boon to heavy-duty network servers.

Tri-Gate transistor is a variant of ‘FinFet’ structure developed by Chenming Hu and his colleagues in late 1990s at University of California, Berkeley. IBM, Samsung and Taiwan Semiconductor Manufacturing Corporation are all working on FinFet. But even this 3-D structure will hit the wall sooner or later. Then perhaps a new technology ‘SPINTRONICS’ will come into picture.

FREQUENCY SPECTRUM ALLOCATION IN 21st Century.

This module gives the need for greater frequency spectrum allocation for Mobile Communication and this has to be fulfilled within 350MHz to 3.5GHz considerable fraction of which is already occupied for TV usage.Hence ways and means have to be devised so that greater portion of this opportunity window is freed for Mobile Applications.

FREQUENCY SPECTRUM ALLOCATION IN 21 st Century.

(Adapted from ’ The Great Spectrum Famine’ by Mitchell Lazarus, Spectrum , November, 2010)

Table 1. The Opportunity Window: The best frequencies for mobile broadband are high enough that the antenna can be made conveniently compact and yet not so high that it fails to penetrate the buildings. This leaves a relatively narrow range of frequencies available for use i.e. from 300MHz to 3.5GHz..

					Best for Mobile	

			Broadband 300Mz to 3.5GHz	
Spectrum Slot	1MHz-10MHz	10MHz-100MHz	100MHz-1GHz	1GHz-10GHz
Size of efficient antenna	75m to 7.5m	7.5m to 0.75m	0.75m to 7.5cm	7.5cm to 7.5mm
	Bounces from the sky Due to ionospheric reflections			Bounces inside the room
	Penetrates buildings and terrain		Reflects from buildings	
	Follows Earth's curvature			Travels along line-of-sight

It is a World Wide Worry:

“The growing popularity of mobile broadband is a global phenomenon, and the laws of physics that limit the frequencies available for it are the same everywhere. So it is no surprise that many other countries are facing the same spectrum famine as the United States.

All regions have large amounts of spectrum dedicated to TV, most of it in the frequency range suitable for wireless broadband. Most countries have targeted a conversion to digital broadcast TV as their best bet obtaining more wireless spectrum-what the European Union calls a digital "dividend".Several have already completed the transition, including Belgium, Denmark, Finland, Germany, Luxembourg, the Netherlands, Norway, Spain, Sweden and Switzerland. Other countries have transition in progress, most scheduled for completion between 2011 and 2013 (China in 2015). The amounts of spectrum harvested by this transition vary from about 100 to 130MHz.

Many countries are also experimenting with spectrum auctions, some with a minimum of regiulation of the technologies, as a way of fostering efficient use."

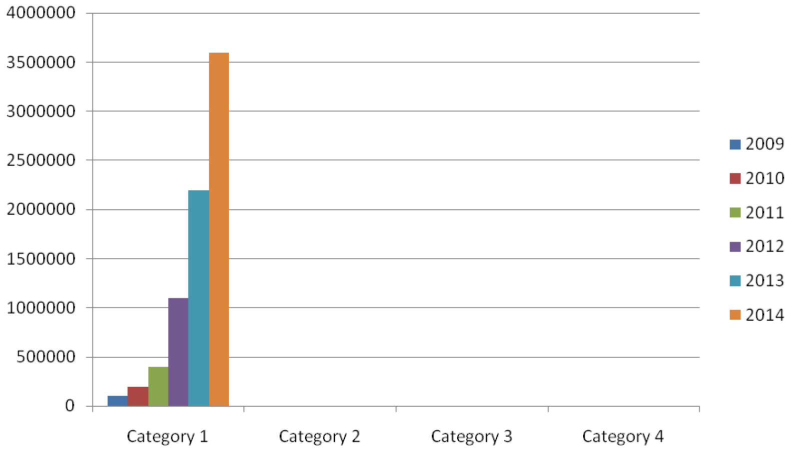


Figure 1.Mobile Broadband Skyrockets Globally. (Source Cisco)

[X-axis Year; Y-axis Terabytes per month.

Terabytes per month includes Voice over Internet Protocol, Gaming, Peer-to Peer networking, Web/ data, Video.

In the year 2014 it is projected that Voice over Internet Protocol (4%), Gaming(6%), Peer-to Peer networking(8%), Web/data(17%), Video(66%).]

Evolution of Mobile Phone Service:

1970s- One transmitter was used to serve the whole city.There were a few dozen voice channels and these channels were shared among the thousands of user in time. Because of time sharing, there were long waits and calls were expensive.

1980s-832 pairs of 30 kHz spaced analog FM channels in the 800-MHz band. These were reused in the cellular layout at different locations across the city to support many thousands of conversations. But service charges were out of reach for the common man.

1990- 1.9-GHz “2G” voice service came in vogue. The service providers used auctioned spectrum for digital transmission which yielded great benefits in terms of quality of voice transmission as well as in terms of efficient use of the spectrum. This was enabled because digital communication has the advantage of error detection and correction as well as it has the advantage of band-width compression.

By the time we entered 21st century, all mobile transmissions had converted to digital transmission resulting in tenfold increase in capacity of the wireless networks.

Transition from Analog to Digital Communication.

The transition from analog to digital mobile phone was smooth and voluntary. Service providers made the change on their own, for their own benefit and on their own time-tables.

The transition from analog TV to digtal TV was slow and painful and in USA it was completed by 2009. But this transition eventually yielded 108MHz of spectrum which were freed for auction to mobile service providers. Now each digtal TV station can accommodate four analog-quality video signals and it has the new options for high definition programming and data services. This results in fivefold improvement in spectrum efficiency.

Ways and means for overcoming this spectrum famine.

1. Identify the least efficient or least critical services and redesign them to use less spectrum. The collective traffic of police, fire-fighters, emergency responders, utility workers, taxi drivers, plumbers, construction crews and

their like can be handled in far less spectrum;
2. Underused mobile satellite bands need to be converted to a primarily terrestrial cellphone-type service;
3. Government keeps large swaths of valuable spectrum. This could be fred for private use.
4. Rapid switch to digital TV is going to yield high dividends.

The upgraded Internet Protocol IPv6 and challenges of expolsively growing INTERNET(Untitled)

The Asia-Pacific Network Information Center, which comprises the fast growing countries India and China was the first RIR to exhaust its allocation of IPv4 addresses on 15th April 2011.Thus it was the first RIR to receive the new block of IPv6 address.

June 6, Wednesday, 2012 was theWorld IPv6 Day<h3>Internet , a trillion times roomier now.

[The Hindu,Thursday, June 7,2012]

Till now we have had IPv4 which allocated 32-bit integer for addresses which permitted 4.3 billion addresses.The Internet Assigned Numbers Authority(IANA) distributes the available internet protocol addresses to the five regional information registries(RIR).The Asia-Pacific Network Information Center, which comprises the fast growing countries India and China was the first RIR to exhaust its allocation of IPv4 addresses on 15th April 2011.Thus it was the first RIR to receive the new block of IPv6 address.

In IPv6 protocol, successor to the IPv4 protocol, there is 128-bit integer addresses to be assigned to hosts or the systems. 128-bit addresses permits 2ˆ128 addresses which amounts to 340 trillion, trillion, trillion $= 3.4 \times 10^{38}$ addresses. This is more than enough for all the new devices which are likely to be connected to the INTERNET in next five years.

IPv6 protocol requires upgraded infrastructure to operate. In phases IPv4 will be eased out and IPv6 will be accordingly introduced so that the whole process is seamless.

Challenges of Video and Mobile access in INTERNET

[Excerpted from the article “The NET was not built keeping Video and Mobile Access in mind”, The Economic Times, Monday, 27th August, 2012, Kolkota edition]

Tom Leighton is the co-founder and chief scientist at Akamai Technologies and he is also a MIT Professor in Computer Science Department.

$1.1 billion servers of Akamai Technologies carry about a third of World’s Web Service and its customers include White House, Facebook, Apple, Twitter, Auch India Infoline and Shopper’s Stop.The following information has been taken from his interview.

Right now in August 2012 there are 35-billion connected devices . This will become 1 trillion in next few years.

Our total population is 7billion as of today. As Internet is penetrating the hinterlands of undeveloped economies, internet connections are exponentially increasing. Also the number of devices per person is increasing The number of devices per person will touch a factor of 10 by 2017.

There is a diversification in the devices. It could be smart phones, tablets or desktops. There is a diversification in their working. Some have CDMA, others are GSM. There is a diversification in their Operating Systems. Some have Android. Apple has its own propriety OS and MicroSoft has its own propriety OS. There is diversification in the formats of Video and Audio transmission such as JPEG, PNG,GIF and TIFF. In audio we have MPEG, Digital Theater System(DTS), Windows Media Audio.

These diversifications create a huge stress on the INTERNET for which INTERNET is not prepared.

Even though Mobile Cellular Net is becoming better , faster and cheaper still it takes very long to get a video download or to get webpage open on cell phones.

Internet is not built for Video Transmissions nor for mobile access. It is not prepared for Social Networking which has come in a big way. This creates network management problems.

On the Internet every different device costs money. The last mile BW is not the same for different devices. Different formats have to be coded/decoded differently. Every added users adds cost.

To remove this cumulative cost build up, we have to bring the diversified messages on a common platform, process it, transmit it, reprocess it and then serve it to the end users at the right bit rates. Algorithm has to be written for these diversified tasks and embedded in the common platform that works for all kinds of devices, across all the formats and across different Band Widths.

Security is the biggest problem and we have to make sure that credit card information are not stolen.

Cellular Phone networks are built for Voice BW in the range 8 to 16 kb/sec. A typical Webpage 100kB. This cannot open the webpage fast. Our goal is instant access on cellular devices. Traditional Internet infrastructure is not capable of handling Video.

In India we want high speed and low cost.

Using applied Maths and Algorithms we are improving the efficiency of our servers. We are also working on the next generation delivery so that most bits come from the Internet and not from our computer.

It is only after these issues are addressed and resolved that we will have a seamless convergence of Video, Audio and Data on the INERNET.,

SSPD_Chapter1_Modern Physics-Foundation of Solid State Physics_Part1

Chapter 1 of SSPD, Part 1 deals with the development of Science through pre-history, ancient history and medieval history upto renaissance.

SOLID STATE PHYSICS

& DEVICES- *the harbinger of the third wave of civilization.*

By Bijay Kumar Sharma

Solid State Physics & Devices- the harbinger of Third Wave of Civilization.

CHAPTER.1.

MODERN PHYSICS- THE FOUNDATION OF SOLID STATE PHYSICS.

INTRODUCTION.

Discoveries or Inventions are made accidently or they are the results of conscious theoretical formulations.

The discovery of electricity through the electric shock experienced by Benjamin Franklin during a ' key and kite' flight in rains*, the discovery of magnetic field surrounding a wire carrying d.c.current during a demonstrative lecture by Michael Faraday and the invention of Bipolar Junction Transistor by William Shockley were some such event which brought a historical transformation of Human Society.

On the other hand the Gravitational Lensing Effect by Massive Bodies as predicted by Einstein based on his General Theory of Relativity** , the prediction of Supernova Explosion and formation of Neutron Star as the end result of Stars more massive than 1.4 times the mass of Sun (MΘ) by S.Chandrashekhar and the prediction of Intermediate Vector Boson as the Boson particle of exchange of weak forces by Abdus Salam and Steven Weinberg were based on conscious theoretical formulation and modeling of the systems concerned.

All our ancient knowledge were based on empiricism hence they were a part of the whole and they never led to a complete understanding of the system concerned.

Our Nomadic Ancestors lived as roots and fruits gatherer in the Savage stage of the society.

With discovery of fire they became scavengers. The cooked meat was easily digestible. Our Ancestors changed from herbivorous to omnivorous. Soon we became toolmakers and hunters. This was the Barbarian stage of pale anthropological society.

About 500,000 years Before the Present (BP), with stone tool making stone age started. We had left our arboreal abode and became cave-dwellers. This was the old stone age or Lower Paleolithic Age.

About 30,000 years B.P. to 10,000 years B.P. Upper Paleolithic Age prevailed.

In Neolithic Age (new stone age), stone axes were made. We started domesticating animals .The need for animal fodder led us to agriculture.

Agriculture triggered the first wave of civilization on the banks of Tigris-Euphrates in Babylon (modern day Iraq), on the banks of Nile river in Egypt, on the banks of Sarswati, Sapt Sindhu and Ganges- Yamuna- Brahmputra in India and on the banks of Yellow River and Yangtze River in China. These ancient civilizations developed and flourished about 3000 years B.P. They were built on the foundation of Neolithic Age.

Agriculture and Pastoral Revolution transformed Human-Beings into active partners of Nature. Through empiricism, foresight, forethought, perseverance and self control we specialized in agriculture and artisanship.

We developed different fields of metallurgy namely smithy, mining and smelting. We developed wheels and ox-cart. We developed sailing ships.

In all these stages and in all these civilizations ,we were dependent on empirical knowledge. We observed ,we learnt and we used it. We never tried to theorize.

In Greek Era, for the first time why, how and whereof questions were raised regarding events, processes and systems. Socrates played the most crucial role in questioning the prevailing dogmas.

In the Greek Era, the prevailing World View was the Ptolemaic World View where our Earth was at the centre of the Universe known as Geo-centric Model and where heart was at the centre of the Human Body known as Cardio-centric Model .

The Greek Era was followed by the Roman Era. King Constantine in 300AD proclaimed Christanity as the State Religion and almost whole of Europe gradually converted to Christanity. On the other hand in 600AD in Arab countries there was the dawn of Islam and Prophet Muhammad was proclaimed 40th Prophet and the last Prophet. Islam spread to whole of Europe where a new resistance started against the spread of Islam in form of Christian Crusades. By 1200AD crusaders managed to push back the Islamic forces back to Arabian Peninsula. Until 15th century Europe remained divided. During this period barbarian Goths, Vandals and Huns swept down on Europe from North and East. They destroyed many fine buildings and works of art that had existed during Roman times. This period was referred to as the dark ages of Medieval Era.During this period knowledge survived only in Monastaries. Progress in Science and Technology came to a stand still.. Whole of Europe was plunged in the dark ages of the Early Medieval Era.

14th Century saw the dawn of the Age of Reasoning. This was the Age of Voltaire.

14th Century also saw the Gutenberg Revolution, the second Information Revolution.***. This century witnessed the Protestant Movement and subsequent rise of Martin Luther as the most significant dissident within the Christian Religion.

The Protestant Movement signaled the Age of Renaissance. The Industrial Revolution of 18th Century catalyzed the Bourgeoisie Democratic Revolution which swept through Europe, USA and Japan. This Revolutionary Movement overthrew the Monarchies and Oligarchies and established the rule of Pluralistic Democracy in the leadership of Bourgeoisie Party. In France this Party was known as Jacobins, in USA it was known as Democratic Party, in Germany it was under the leadership of Junkers and in Japan it was under the leadership of Meiji Oligarchs.

[**Benjamin Franklin made a systematic and scientific* *investigation of lightening during the second half of 18th Century through his legendary ‘Key and Kite’experiment. In a Pennsylvania thunderstorm in 1752, he flew a kite with a key tied to the string just beside the kite. The electric charges induced on the thunderclouds were attracted by the key and provided a short circuit path to the ground through the body of Franklin. In the process he received electric shocks which could have proved fatal. But this electric shock established the nature of electricity as well as the charge of electricity which in this case was negative. Thus electricity’s discovery was made and this was to play the same role in human society as the discovery of fire was to play in the hominid society when hominid society was fruit and root gatherer as well as scavenger about a million years before.*

*** The General Theory of Relativity predicts that the Space-Time Fabric is flat in absence of mass but presence of massive celestial bodies causes a curvature in this fabric. Hence it was predicted that a star behind the sun during a total solar eclipse will be seen shifted from its actual position by 1.7second of arc. Sir Arthur Stanley Eddington (1882-1944) went to Canary Islands and made actual observation of a star behind Sun during Total Solar Eclipse on 29 th May,1919. Next the same star was observed in absence of Sun. Near about 1.7second of an arc displacement was observed in the apparent and actual position of the star. This vindicated Einstein and his General Theory of Relativity.*

In recent days the gravitational lensing effect of distant quasers has lent further support to the General Theory of Relativity.

The General Theory of Relativity gives the theoretical basis of precession of the perihelion of the planet mercury.

The General Theory of Relativity predicts Red Shift of light coming from massive body. This too has been correctly verified by observation.

*** The First Information Revolution was ushered in by the invention of Phoenician Alphabets which was the mother of all European Languages. The invention of the Rotating Press was the Gutenberg Revolution and the second Information Revolution.

The Computer Revolution is the Third Information Revolution.]

A question which automatically arises at this point is that why the third World Countries did not witness Bourgeoisie Democratic Revolution. No general theory can be offered to explain this aberration but in context with India an explanation can be given.

India lies in Tropical Climatic Zone where the year round the weather is favorable for agriculture. We are able to grow most of our food and fodder requirements. Hence any form of Trade , land or mercantile trade, never held much significance for us. Whereas Foreign Trade is an important component of total Gross National Product of European Nations. Foreign Trade implies Mercantile Trade and Mercantile Trade means long Oceanic Voyages. Through these Oceanic Voyages Christopher Columbus reached the New World and Vascodagama discovered India, the golden bird, and Francis Drake navigated the whole World circumferentially.

The long Oceanic voyages compelled European Nations to invest heavily in Navigation Sciences in particular and in Science & Technology in general. This is how sextant and astrolabe were invented. Even after the Old Democratic Revolution, the healthy trend of patronizing and financing Science & Technology was maintained. As a result the credit of all the major Discoveries and Inventions went to Western and Japanese Scientists except for six which came to Indian Scientists.

These Scientific breakthroughs brought the European Nations on the path of Colonization through the Mercantile Trades which was a logical end.

Colonization led to a highly skewed economic development. The First World countries achieved a much higher level of prosperity than what they deserved whereas the Third World countries were pushed deeper and deeper into the mire of poverty and deprivation. The chasm between West and East has been widening ever since.

Today the responsibility lies on the Intellectuals and Scientists of Third World Countries to chalk out an alternative paradigm of development whereby there is equitable and just development of all geographical regions of our Globe.

Our readers must be anxious to know as to what are those four discoveries and Inventions for which Indian scientists are credited?

First is the fundamental work of Ramanujam on number theory and higher Mathematics [Appendix I]. Second is the work of Meghnath Saha (1893 -1956) on heat and molecular Physics. His ionization equation, Saha-Langmuir Equation, plays an important role in stellar astro-physcics [Appendix II].Third is the discovery of Bose-Einstein Statistics and the discovery of Bosons by Satyendra Nath Bose **[Appendix III]** .Fourth is the discovery of Raman Effect by C.V. Raman [Appendix IV]. This fetched us a Nobel Prize.

Fifth is the extensive entomological work done on Mosquitoes by Ronald Ross **[Appendix V]** .This also fetched a Nobel Award.

Sixth is the work done by G.N. Ramchandran (1922-2001) on molecular Biophysics. He is considered the father of this field[Appendix VI]. After Independence our five major achievements have been in the area of green revolution, space research, nuclear energy, super- conductivity and CSIR (Council of Scientific and Indistrial Research) into a locomotive of Science and Technology advancements.Here we befittingly mention the works done by Indian Scientists working abroad notably in USA. First is the discovery of Chandrshekher Limit S. Chandra Shekher in the area of Evolution of Main Sequence Stars[Appendix VII]. Second is the work of Har Gobind Khurana(1922-till date) in the field of genetics and its function in protein synthesis[Appendix VIII].

Third is the work done again in Genetics by Venktaraman Ramakrishna. Therte are DNA molecules inside every cell of all organisms. These molecules contain the blue-print for how a human being or a plant or a bacterium looks and functions. The blue prints of the DNA molecules become transformed in the living matter through the ribosomes. Based upon the information in DNA , ribosomes create protein: oxygen transporting haemoglobins, antibodies of the immune system, hormones succh as insulin, collagen of the skin, or enzymes. There are tens of thousands of protein in the body and they all have different forms and functions. 2009 Nobel Prize for Chemistry Awards Research in one of life's most important processes: the ribosomes translation of DNA information into life [Appendix IX]. Here it will be appropriate to mention invention of Wireles by Jagdish Chandra Bose [Appendix X]. Though credit was accorded to Marconi, lately the mistalke has been corrected and in 1997 IEEE (Institution of Electronics and Electrical Engineering) has acknowledged the pioneering role of Dr. Jagdish Chandra Bose in the field of Wireless and Radio.After this digression we return to our main concern namely Solid State Physics and Devices.

The present development and widespread use of Electronics in every walk of life is primarily due to Integrated Circuit Technology and due to the continuous evolution of the level of Integration as stated by Moore’s Law. Gordon Moore **[Appendix XI]** is a founder of Intel Corporation which caters to 80% of microprocessors market.

Gordon Moore had predicted at the inception of Integrated Circuit Technology that the packing density of CMOS devices will double every 18 months i.e. every one and a half year . That precisely has been followed till date.

Integrated Circuit technology was simultaneously invented by Jack Kilby **[Appendix XII]** of Texas Instrument and Robert Noyce **[Apendix XIII]** of Fairchild in 1959. Eventually Jack Kilby was awarded Nobel Prize in 2000 for the same. I. C. Technology **[50 years Journey of IC Technology-**

Prologue] started its journey in 1961 when the first RTL logic was commercially marketed. At that time there were only twenty transistors on the Chip. Today in 2005 there are 1 billion transistors on Pentium “Extreme Edition” 840 Dual Core Chip with a clock rate 3.2GHz.

This hectic pace of microminiaturization has been maintained solely due to commensurate development in Solid State Device Technology **[Apendix I, Table I.1, of 50 years I.C. Journey]** .

The development in Solid State Device Technology is the result of the theoretical understanding of Solid State of Matter. And the theoretical understanding has been made possible due to Quantum Mechanics.

So it is imperative that we start our text on Solid State Physics and Devices with a good qualitative and quantitative understanding of Quantum Mechanics.

1.1. MAX PLANCK AND QUANTUM THEORY.

In 1860 Maxwell **[Appendix XIV]** propounded the four classical equations describing Electro-Magnetic Waves namely:

1. Divergence of Electric Flux Density = Charge Density,
2. Curl of Magnetic Field Intensity = Electric

Current Surface Density + Displacement Current Density;
3. Curl of Electric Field Intensity = -Rate of change of Magnetic Flux Density;
4. Divergence of Magnetic Field Intensity = 0;

Almost simultaneously Heinrich Hertz **[ibid]** generated Electromagnetic Waves and detected the same through a loop antenna. This particular experiment paved the way for the invention of Radio for which both Marconi **[Appendix XV]** and J.C.Bose **[Appendix X]]** are credited.

The theoretical explanation of Black Body Radiation kept eluding the Scientists at the close of 19th Century. Rayleigh-Jean formulation could explain the longer wavelength spectral density distribution but on the shorter wavelength side it led to Ultra-Violet catastrophe. On the other hand Wien Formulation could explain the shorter wavelength spectral density distribution but failed to fit the remaining portion of the distribution curve. The Black Body radiation curve is described in Figure (1.1).

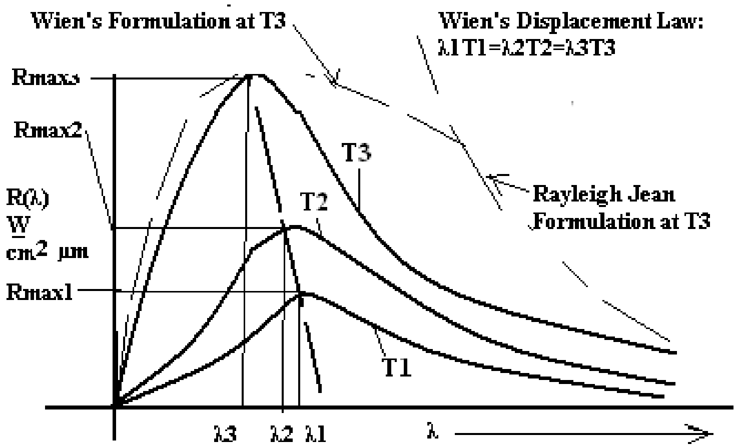


Figure 1.1.The spectral distribution of Black Body Radiation at different temperatures: T1<T2<T3.

Fig.1.1. The Spectral Density Distribution of Black Body Radiation at different Temperatures: T 1 > T 2

At this time of hopelessness, a simple assumption by Max Planck enabled him to arrive at the correct theoretical understanding of Black Body Radiation Distribution. His assumption was that matter is composed of oscillators. Every oscillator has a characteristic frequency ν_0 and that oscillator will absorb or radiate energy packets of $h\nu_0$, 2 $h\nu_0$,3 $h\nu_0$,…… only. Here h is the Planck’s Constant$=6.62\times10^{-34}$ J-seconds. The essence of this assumption is the quantum nature of Electromagnetic Waves which was the basic proposition of Newton also who talked about the corpuscular nature of light which is also Electromagnetic Waves. According to this quantum nature , Electromagnetic Waves of frequency ν

always contains an energy packet of $h\nu$ Joules. Therefore if a monochromatic light source of frequency ν is emitting energy at the rate of P watts then $P/(h\nu)$ number of energy packets are being emitted per second. More appropriately $P/(h\nu)$ number of quanta or photons are being emitted by the source per second. If the power of this source is increased then the number of photons per second will increase but the energy packet per photon will remain constant.

Thus quite unwittingly Max Planck gave birth to the Quantum Theory of Electromagnetic Waves..

Einstein used this Quantum Theory to explain Photo-Electric Effect.

Wien's Displacement Law [Appendix XVI] states :

λm(wavelength of maximum intensity) = **Ω/T**;

Classical Physics has no analytical value of Ω but Quantum Physics correctly predicts it to be : Ω = **ch/(4.9651k)** = **3×10-3 mK**;

Substituting this constant in Wien's Displacement Law: $\lambda_m = (3\times10^{-3}/T)$ m;

At 3 K, $\lambda_m = 0.1$ cm. The Cosmic Microwave Background Radiation(CBR), which we will deal with in Section 1.13 , has exactly this 3K blackbody radiation characteristics.

Stefan Boltzman’s Law [ibid] states:

R(Watt per m2) = e.σ.Tˆ4 ;

σ = Stefan’s Constant;

e= emissivity (depends on the nature of the radiating surface);

e=1 means a black body radiator. A black body is a perfect absorber. It is a perfect radiator also.

e= 0 means a perfect reflector. It does not absorb and hence does not radiate.

This law is also known as Fourth Power Law. T is the absolute temperature of the cavity walls.

From Quantum Theory of Black-Body Radiation:

u(total energy density of the radiation in a cavity) = a.T ˆ4

Classical Physics has no analytical value of σ but Quantum Mechanics gives the value as:

a(Universal Constant) = $8\pi^5k^4/(15c^2h^3)$. This value conforms to the experimental values.

e (emissivity) = 0 to 1.

σ (Stefan’s Constant) = ac/4 = 5.670×10^{-8} W/

(m.K4);

end of part1

SSPD Chapter 1_Part 1_Appendices of Disoverers and Inventors
This module gives the appendices from Appendix I to Appendix XV

Appendix I.

S. Ramanujan

The great Indian mathematician, S. Ramanujan has left the sign of his brilliance in various fields of mathematics like Algebra, Geometry, Trigonometry, Calculus, Number theory etc. throughout his entire life. He has also made some extraordinary contributions to the fields like Hypergeometric series, Elliptic functions, Prime numbers, Bernoulli`s numbers, Divergent series, Continued fractions, Elliptic Modular equations, Highly Composite numbers, Riemann Zeta functions, Partition of numbers, Mock-Theta functions etc. Actually, apart from a few elementary ones, most of the contributions of S. Ramanujan belong to a higher realm of mathematics that is often referred to as "Higher Mathematics". In fact, one can find it quite difficult to understand S. Ramanujan`s mathematics if he does not have the basic foundation in various mathematical subjects.

Appendix II

Meghnad Saha (October 6, 1893 – February 16,

1956) was an Indian astrophysicist best known for his development of the Saha equation, used to describe chemical and physical conditions in stars.

The **Saha ionization equation**, also known as the **Saha-Langmuir equation**, was developed by the famous Indian astrophysicist Megh Nad Saha in 1920, and later (1923) by Irving Langmuir. One of the important applications of the equation was in explaining the spectral classification of stars. The equation is a result of combining ideas of quantum mechanics and statistical mechanics.

"Meghnad Saha’s ionization equation (c. 1920), which opened the door to stellar astrophysics” was one of the top ten achievements of 20th century Indian science [and] could be considered in the Nobel Prize class." - Jayant Narlikar

"The impetus given to astrophysics by Saha’s work can scarcely be overestimated, as nearly all later progress in this field has been influenced by it and much of the subsequent work has the character of refinements of Saha’s ideas." - S. Rosseland[4]

"He (Saha) was extremely simple, almost austere, in his habits and personal needs. Outwardly, he sometimes gave an impression of being remote, matter of fact, and even harsh, but once the outer shell was broken, one invariably found in him a person of extreme warmth, deep humanity, sympathy and understanding; and though almost altogether unmindful of his own personal comforts, he was extremely solicitous in the case of others. It was not in his nature to placate others. He was a man of undaunted spirit, resolute determination, untiring energy and dedication." - D. S. Kothari

Appendix III

Satyendra Nath Bose

(1 January 1894 – 4 February 1974), FRS, was an Indian physicist, specializing in mathematical physics. He is best known for his work on quantum mechanics in the early 1920s, providing the foundation for Bose-Einstein statistics and the theory of the Bose-Einstein condensate. He is honoured as the namesake of the boson.

Although more than one Nobel Prize was awarded for research related to the concepts of the boson, Bose-Einstein statistics and Bose-Einstein condensate—the latest being the 2001 Nobel Prize in Physics, which was given for advancing the theory of Bose-Einstein condensates—Bose himself was never awarded the Nobel Prize. Among his other talents, Bose spoke several languages and could also play the *Esraj*, a musical instrument similar to a violin.

In his book, *The Scientific Edge*, the noted physicist Jayant Narlikar observed:

> “ S. N. Bose’s work on particle statistics (c. 1922), which clarified the behaviour of photons (the particles of light in an enclosure) and opened the door to new ideas on statistics of Microsystems that obey the rules of quantum theory, was one of the top ten achievements of 20th century Indian science and could be considered in the Nobel Prize class. [1] ”

Bose's ideas were afterwards well received in the world of physics, and he was granted leave from the University of Dhaka to travel to Europe in 1924. He spent a year in France and worked with Marie Curie, and met several other well-known scientists. He then spent another year abroad, working with

Einstein in Berlin. Upon his return to Dhaka, he was made a professor in 1926. He did not have a doctorate, and so ordinarily he would not be qualified for the post, but Einstein recommended him. His work ranged from X-ray crystallography to unified field theories. He also published an equation of state for real gases with Megh Nad Saha.

Apart from physics he did some research in biochemistry and literature (Bengali, English). He made deep studies in chemistry, geology, zoology, anthropology, engineering and other sciences. Being an Indian of Bengali descent, he devoted a lot of time to promoting Bengali as a teaching language, translating scientific papers into it, and promoting the development of the region.

In 1944 Bose was elected General President of the Indian Science Congress.

In 1958 he became a Fellow of the Royal Society.

Appendix IV.

C.V. Raman.

(7 November 1888 – 21 November 1970) was an Indianphysicist and Nobel laureate in physics recognised for his work on the molecular scattering of light and for the discovery of the Raman effect, which is named after him.

In 1948 Raman, through studying the spectroscopic behavior of crystals, approached in a new manner fundamental problems of crystal dynamics. He dealt with the structure and properties of diamond, the structure and optical behavior of numerous iridescent substances (labradorite, pearly feldspar, agate, opal, and pearls). Among his other interests were the optics of colloids, electrical and magnetic anisotropy, and the physiology of human vision.

Raman was honoured with a large number of honorary doctorates and memberships of scientific societies. He was elected a Fellow of the Royal Society early in his career (1924) and knighted in 1929. In 1930 he won the Nobel Prize in Physics. In 1954 he was awarded the Bharat Ratna. He was also awarded the Lenin Peace Prize in 1957.

India celebrates National Science Day on 28 February of every year to commemorate the discovery of the Raman effect in 1928.

Appendix V.

Ronald Ross

(13 May 1857 – 16 September 1932) was a Scottish physician.

Ross commenced his study of medicine at St. Bartholomew's Hospital in London on October 29, 1875. He passed his final examination in 1880 and qualified as MRCS and LSA. He joined the Indian Medical Service in 1881. His first posting was in Madras.

Ross received many honours in addition to the Nobel Prize, and was given Honorary Membership of learned societies of most countries of Europe, and of many other continents. He got an honorary M.D. degree in Stockholm in 1910 at the centenary celebration of the Caroline Institute and his 1923 autobiography *Memoirs, Etc.* was awarded that year's James Tait Black Memorial Prize. Whilst his vivacity and single-minded search for truth caused friction with some people, he enjoyed a vast circle of friends in Europe, Asia and the United States who respected him for his personality as well as for his genius.

Ross married Rosa Bessie Bloxam in 1889. They had two sons, Ronald and Charles, and two daughters, Dorothy and Sylvia. His wife died in 1931. Ross survived her until a year later, when he died after a long illness, at the Ross Institute, London, in 1932.

In India Ross is remembered with great respect. Because of his relentless work on malaria, the deadly epidemic which used to claim thousands of lives every year could be successfully controlled. There are roads named after him in many Indian towns and cities. In Calcutta the road linking Presidency General Hospital with Kidderpore Road has been renamed after him as Sir Ronald Ross

Sarani. Earlier this road was known as Hospital Road. In Hyderabad, the famous Quarantine (Koranti) hospital is named as Sir Ronald Ross Institute of Tropical and Communicable Diseases in recognition to his services in the field of tropical diseases.

Appendix VI

G.N. Ramachandran

Gopalasamudram Narayana Iyer Ramachandran, or G.N. Ramachandran,

(8 October 1922 - 7 April 2001) is widely acknowledged as one of the most important Indian scientists of the 20th century, best known for his work that led to his creation of the Ramachandran plot for understanding peptide structure. He was the first to propose a triple-helical model for the structure of collagen. He also made other major contributions in biology and physics.

Professor Ramachandran was born in the town of Ernakulam, Kerala, India. He joined the Indian Institute of Science, Bangalore in 1942 in the Electrical Engineering Department. Gopalasamudram, the native place of his forefathers, is a village in the old Tirunelveli District of Madras Presidency. Quickly realizing his interest in physics, he switched to the Department of Physics to complete his master's and doctoral thesis under the supervision of Nobel laureate Sir C. V. Raman. In 1942 he received a master's degree in physics from Madras University with his thesis submitted from Bangalaore (he did not attend any Madras college at that time). He subsequently received his D.Sc. degree in 1947.[1] Here he mostly studied crystal physics and crystal optics. During his studies he created an X-ray focusing mirror for the X-ray microscope. The resulting field of crystal topography is used extensively in studies involving crystal growth and solid-state reactivity.

Ramachandran then spent two years (1947–1949) at the Cavendish Laboratory in Cambridge, where he earned his Ph.D. for 'studies on X-ray diffuse scattering and its application to determination of elastic constants' under the direction of Professor Wiiliam Alfred Wooster, popularly known as W.A. Wooster, a leading crystollagraphy expert in the world.

Appendix.VII.

S. Chandrashekhar.

(S. Chandrashekhar. By Kameshwar Wali, Imperial College Press , 1997)

Subrehmanyam Chandrashekher was an austere, diligent and patient scientist and a very humane person. He was born in 1910. In 1930 he sailed to Cambridge University, England. While in the ship , he worked out the fate of collapsing stars. Stars less than 1.4Mɵ after burning out their nuclear fuel collapses into white dwarf where electrons are tightly packed to their spatial limit and develop degeneracy pressure which halts further collapse. But stars heavier than 1.4Mɵ overcome this degeneracy pressure and collapse into neutron star. Chandrashekher had correctly predicted that Supernova remnant will be a neutron star.

Ralph Fowler used Fermi-Dirac Satistics to explain the degeneracy pressure of electrons which helps prevent further collapse of White Dwarf after the fusion reaction fuel is exhausted. Fowler calculated that for a given composition, matter density of white dwarf was proportional to square of the mass. But this led to the impossible propostion that if the mass of a dying star (where all the hydrogen has ben used up) was 1.4Mo (where Mo is the mass of our Sun) or greater, then density tends to become infinity.

S. Chandrashekher used Special Theory of Relativity to explain the state of matter if the mass of the dying star exceeded 1.4Mo and concluded that if star is heavier than 1.4Mo then super-nova explosion takes place and the star collapses into neutron star and the degeneracy pressure of neutrons prevents the further collapse of the netron star.

In 1935, he presented his finding to Royal Astronomical Society but his own mentor and research guide Arthur Eddington denounced as Stellar Buffonery. S. Chandrashekher was dejected and crest fallen. He moved to an Academic Position in Chicago University in the year 1937. In 1939 he published his monograph on Stellar Structure including his findings on neutron star.

In 1939 J. Robert Oppenheimer and his student Hartland Snyder wrote a joint paper "On Continued Gravitational Collapse". In this paper General Theory of Relativity was applied to explain the state of matter and it was given the name of Black Hole. It was conjectured that if star is heavier than 3Mo then even the neutron degeneracy does not stop the collapse. Now matter is a newer form called Black Hole which leads to countering the gravitational collapse.

It was clearly shown that there was no such thing as peaceful equilibrium. The criteria of equilibrium was continuously changing. We are essentially in a violent world.

Fritz Zwicky observed supernove explosions and identified them as the explosions which dying stars experience when Mass of the star exceeds 1.4Mo. When mass is below 1.4Mo then it experiences only nova explosion. Supernova exploding stars were regarded as standard candle for measuring the distances of galaxies. Fritz Zwicky also recognized the dark matter whose gravitation dominates the dynamics of large scale structure.

Arthur Edington and Edward Arthur Milne rejected Chandrashekeher's conjecture of 1.4Mo. Similarly Edwin Huble and Walter Baade a close colleague of Fritz Zwicky rejected the conjecture of Zwicky that of a violent Universe. But finally both S.Chandrashekher and Zwicky were vindicated.

In 40's S. Chandrashekher offered a course on elementary particles for Yerkes University for which he used to drive down 100miles from Chicago University. But his class in this course consisted of only two students. He considered this worth the effort and eventually in 1957 his belief was vindicated. His both the students Chen Ning Yang & Tsung Dao Lee shared the Nobel Prize in Physics in 1957 for their penetrating investigations in the parity laws. The Universe is just wee bit asymmetrical as a result there are instances of parity law violations. This has far reaching consequences for Elementary Particle Physics.

1960's saw further vindication of S. Chandrashekher when in 1967 first PULSARS were detected. Discovery of PULSARS made Arthur Eddington eat a humble pie.

In 1961 S. Chandrashekher started work on General

Theory of Relativity. In 1982 his authoritative treatise on Black Holes was published. 1983 saw his ultimate vindication when he shared Nobel Prize in Physics with William A.Fowler. S. Chandrashekher won his laurels for his original doctoral dissertation in 1930 which established Chandrashekher Limit 1.4MΘ. Fowler won his award for his theory of Nucleosynthesis of elements.

In 1999, NASA named its Chandra X-ray Orbiting Observatory after him.

Appendix VIII

Har Gobind Khurana

Har Gobind Khorana, or **Hargobind Khorana** (born January 9, 1922) is an Indian American molecular biologist. He was awarded the Nobel Prize in Physiology or Medicine (shared with Robert W. Holley and Marshall Warren Nirenberg) in 1968 for his work on the interpretation of the genetic code and its function in protein synthesis. Khorana and Nirenberg were also awarded the Louisa Gross Horwitz Prize from Columbia University in the same year. He became a naturalized citizen of the United States in 1966, and subsequently received the National Medal of Science. He currently lives in Cambridge, Massachusetts, United States serving as MIT's Alfred P. Sloan Professor of Biology and Chemistry, Emeritus.

Appendix IX

The Royal Swedish Academy of Sciences has decided to award the **2009 Nobel Prize in Chemistry jointly to**

- **Venkatraman Ramakrishnan**, MRC Laboratory of Molecular Biology, Cambridge, United Kingdom

- Thomas A. Steitz, Yale University, New Haven, CT, USA

- Ada E. Yonath, Weizmann Institute of Science, Rehovot, Israel

"for studies of the structure and function of the ribosome"

This year's Nobel Prize for Chemistry awards research into one of life's most important processes: the ribosome's translation of DNA information into life. Ribosomes produce proteins, which in turn control the chemistry in all living organisms. There are DNA molecules inside every cell of all organisms. These molecules contain the blueprints for how a human being, a plant or a bacterium, looks and functions.

The blueprints of the DNA molecules become transformed into living matter through the ribosomes. Based upon the information in DNA, ribosomes create proteins: oxygen-transporting haemoglobin, antibodies of the immune system, hormones such as insulin, the collagen of the skin, or enzymes. There are tens of thousands of proteins in the body and they all have different forms and functions.

The knowledge about the ribosome's inner workings is put to practical use in many of today's antibiotics which cure diseases by blocking the function of bacterial ribosomes. Bacteria cannot survive without functional ribosomes, so ribosomes are an important target for new antibiotics.

Appendix X

Jagdish Chandra Bose

(November 30, 1858 – November 23, 1937) born in a [[Bengali people|Bengali] HinduKayasth family was a polymath: a physicist, biologist, botanist, archaeologist, and writer of science fiction.[1] He pioneered the investigation of radio and microwaveoptics, made very significant contributions to plant science, and laid the foundations of experimentalscience in the Indian subcontinent.[2] He is considered one of the fathers of radio science,[3] and is also considered the father of Bengali science fiction. He was the first person from the Indian subcontinent to get a US patent, in 1904.

He made remarkable progress in his research of remote wireless signaling and was the first to use semiconductor junctions to detect radio signals. However, instead of trying to gain commercial benefit from this invention Bose made his inventions public in order to allow others to develop on his research. Subsequently, he made some pioneering discoveries in plant physiology. He used his own invention, the crescograph, to measure plant response to various stimuli, and thereby scientifically proved parallelism between animal and plant tissues. Although Bose filed for a patent for one of his inventions due to peer pressure, his reluctance to any form of patenting was well known. He is being recognised for many of his contributions to modern science.

Bose's first scientific paper, "On polarisation of electric rays by double-refracting crystals" was communicated to the Asiatic Society of Bengal in May 1895, within a year of Lodge's paper. His second paper was communicated to the Royal Society of London by Lord Rayleigh in October 1895. In December 1895, the London journal the *Electrician* (Vol 36) published Bose's paper, "On a new electro-polariscope". At that time, the word 'coherer', coined by Lodge, was used in the English-speaking world for Hertzian wave receivers or detectors. The *Electrician* readily commented on Bose's coherer. (December 1895). *The Englishman* (18 January 1896) quoted from the *Electrician* and commented as follows:

”Should Professor Bose succeed in perfecting and patenting his ‘Coherer’, we may in time see the whole system of coast lighting throughout the navigable world revolutionised by a Bengali scientist working single handed in our Presidency College Laboratory.”

Bose planned to “perfect his coherer” but never thought of patenting it.[13]

His next contribution to science was in plant physiology. He forwarded a theory for the ascent of sap in plants in 1927, his theory contributed to the vital theory of ascent of sap. According to his theory, electromechanical pulsations of living cells were responsible for the ascent of sap in plants.

Appendix XI

Gordon Moore

Gordon Moore along with Robert Noyce established INTEL (Integrated Electronics) . Along the way they were joined by Andy Grove. INTEL went on to produce microprocessor chips [This story is given in ‘Unfolding of Third Information Revolution’] and eventually INTEL captured 80% of the microprocessor market.

Moore's law is the empirical observation that at our rate of technological development, the complexity of an integrated circuit, with respect to minimum component cost will double in about 24 months.

Moore had heard Douglas Engelbart's similar observation possibly in 1965. Engelbart, a co-inventor of today's mechanical computer mouse, believed that the ongoing improvement of integrated circuits would eventually make interactive computing feasible.

Appendix XII

Jack Kilby

6/22/2005 12:03:05 PM, by Eric Bangeman

Is it too much of an exaggeration to say that if it wasn't for integrated circuits, none of us would be doing this right now? Jack Kilby, who is credited with the invention of the integrated circuit, died this week at the age of 81 after a brief battle with cancer.

The invention of the integrated circuit came about as the result of his being left nearly alone at Texas Instruments not long after he was hired in 1958. The plant he was working at shut down for a summer holiday, but as a new hire, he didn't have any vacation time coming. As a result, he was able to spend some time pondering how to further shrink the size of the transistor. From Kilby's recollection of the invention of the integrated circuit:

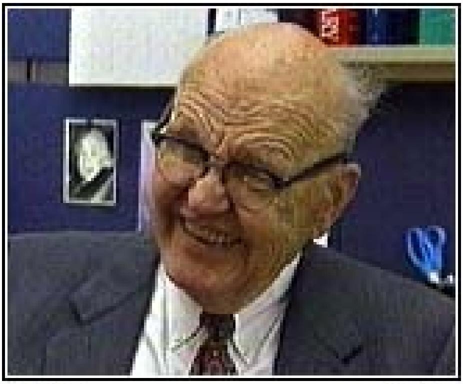
"Further thought led me to the conclusion that semiconductors were all that were really required - that resistors and capacitors, in particular, could be made from the same material as the active devices.

"I also realized that, since all of the components could be made of a single material, they could also be made in situ, interconnected to form a complete circuit. I then quickly sketched a proposed design for a flip-flop using these components. Resistors were provided by bulk effect in the silicon, and capacitors by p-n junctions."

Texas Instruments made the invention public in March 1959, and received the first patent for the device. Robert Noyce, who eventually cofounded Intel with Gordon Moore, was also working on the integrated circuit around the same time at Fairfield Semiconductor and is credited with making mass production of the integrated circuit feasible.

In addition to also receiving over 60 patents, Kilby was a 2000 Nobel Prize in Physics recipient for his invention of the integrated circuit.

'Humble giant' hailed for inventing integrated circuit



Kilby (CNN)

Computer maker launches observance of 1958 development

September 9, 1997Web posted at: 9:12 p.m. EDT (0112 GMT)

DALLAS (CNN) -- He's a tall, non-assuming man who speaks in a modest voice that resounds with a gentle Texas twang. It would be hard to pick him out as one of the most important men of this century. But Jack Kilby is.

Almost 40 years ago this week, Kilby came into work on a day off and put together an invention that came to be called the integrated circuit. It was one of the most significant developments of the 20th century.

CNN's Charles Zewe reports 2 min., 30 sec. VXtreme streaming video

"It looked like a useful idea and one that was worth pursuing," the Dallas engineer said in a recent CNN interview.

It certainly was.

Kilby's integrated circuit became to the information age what the internal combustion engine was to the automobile.

Almost no aspect of modern life has been untouched by the integrated circuit. It runs everything from fax machines and computers to telephones and televisions.

"It affects you every day of your life .. probably 10 times every hour," said Thomas Engibous, Texas Instruments president.

Texas Instruments in Dallas has launched a year-long observance of Kilby's September 12, 1958, discovery.

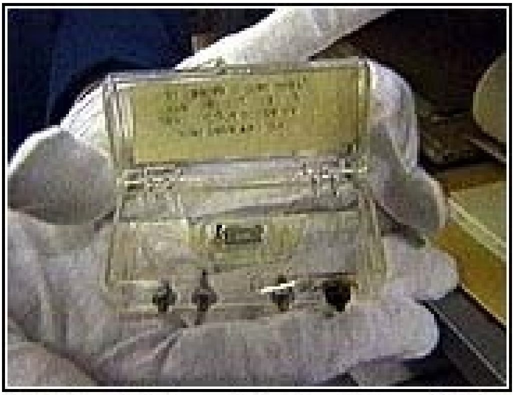
The company has named a new $154 million digital chip research center after him and given his foundation $1 million to honor scientists and inventors.

Known as the "humble giant" around Texas Instruments' headquarters for his height and soft-spoken modesty, Kilby claims to be computer illiterate.

Kilby says he is astounded by what his invention has done.

"What we didn't realize then was that the integrated circuit would reduce the cost of electronic functions by a factor of a million to one," Kilby said.

"Nothing had ever done that for anything before," he said.



The first integrated circuit (CNN)

Doing things that no one had done before seems to come naturally to Kilby, because years after inventing the integrated circuit he went on to co-invent the portable calculator.

His original chip had only a single transistor, and was about the size of a small finger. Today, a chip smaller than a dime can hold 2 million transistors. And a thumbnail-sized area on a chip wafer can hold 125 million transistors.

Asked if there is anything he regrets about the technology that has driven space exploration, revolutionized military weaponry and spawned millions of consumer products, Kilby says there is one regret that springs to mind.

Electronic greeting cards that deliver annoying messages.

Correspondent Charles Zewe contributed to this report.

Appendix XIII

Robert Noyce

Robert Noyce was with FairChild. While he was in FairChild in 1959 , he invented Planar Integrated Circuits almost simultaneously with Jack Kilby of Texas Instrument. Eventually he left FairChild and along with Gordon Moore established INTEL which was eventually joined by Andy Grove. INTEL went on to capture 80% of microprocessor market. Robert Noyce is credited along with Jack Kilby for the invention of Integrated Circuits but since he died in 90’s he could not share the Nobel Prize with Jack Kilby. Nobel Prize is not given posthumously.

APPENDIX(XIV)

Wave Equation and Plane Propogating Waves.

A short chronology of related events.

Galileo (1564-1642) invented telescope, used it to study the celestial bodies and established Heliocentric World View as propounded by Copernicus quite contrary to Geocentric World View held sacred by Ptolemaic adherents and by the Roman Catholic Church. Galileo was excommunicated and put under house arrest for life. Only recently in 1980's excommunication was lifted and Galileo was accepted as a good Christian.

Newton(1642-1727) wrote the Principia, discovered light dispersion by prism and gave the theory of the corpuscular nature of light.

Euler, Young, Fresnel & Foucalt (1819-1868) they all established the wave nature of light through interference and diffraction experiment.

James Clark Maxwell(1864) gave the celebrated four equations which unify Electric and Magnetic Fields.

Heinrich Rudolf Hertz(1857-1894) verified the existence of long electro-magnetic waves. His transmitter was an oscullatory discharge across a spark gap(a form of oscillating electric dipole). For detecting the waves he used an open loop of wire with a brass knob on one end and a fine copper point on the other. A small spark visible between the two ends marked the detection of an incident electromagnetic waves. He focused it , he determined its polarizations, reflected and refracted it, caused interference and produced standing wave and measured its wavelength of the order of 1 m. After his experiment he wrote the following:

I have succeeded in producing distinct rays of electric force, and in carrying out with them the elementary experiments which are commonly performed with light and radiant heat……… We may perhaps further designate them as rays of light of very great wavelength. Theexperiments described appear to me , at any rate, eminently adated to remove any doubt as to the identity of light, radiant heat, and electromagnetic wave motion. [Reference: Optics, Eugene Hecht, Addison-Wesley, 1998, 3rd edition].

Max Planck(1858-1947) was able to explain Black Body Radiation in 1900 by invoking the quantum nature of light though he himself was not comfortable with this new conceptualization.

Einstein(1879-1955)- In 1905 The Special Theory of Relativity was proposed which stated the guage invariance of the speed of light.

1. In 1910 The Photo-Electric Effect was discovered
2. In 1915 The General Theory of Relativity was proposed which predicted Gravitational Lensing Effect.

Michelson & Morley Experiment performed in 1887 which established the guage invariance of the speed of light.

Prince of France- de Broglie (1924) proposed duality theory and matter wave.

Davisson & Germer(1927)- electron diffraction by nickel foils.

Divergence D $= \rho$ (Gauss's Law)

Divergence B $= 0$ (No magnetic monopole)

Curl E $= -\partial B/\partial t$ (Faraday's Law and Lenz's Law)

Curl H $= \partial D/\partial t$ (Displacement Current) $+$ J(electronic current)

Appendix XV

Marchese **Guglielmo Marconi**(25 April 1874– 20 July 1937) was an Italian inventor, best known for his development of a radiotelegraph system, which served as the foundation for the establishment of numerous affiliated companies worldwide. He shared the 1909 Nobel Prize in Physics with Karl Ferdinand Braun, "in recognition of their contributions to the development of wireless telegraphy".

During his early years, Marconi had an interest in science and electricity. One of the scientific developments during this era came from Heinrich Hertz, who, beginning in 1888, demonstrated that one could produce and detect electromagnetic radiation—now generally known as "radio waves", at the time more commonly called "Hertzian waves" or "aetheric waves". Hertz's death in 1894 brought published reviews of his earlier discoveries, and a renewed interest on the part of Marconi. He was permitted to briefly study the subject under Augusto Righi, a University of Bologna physicist and neighbor of Marconi who had done research on Hertz's work. Righi had a subscription to The Electrician where Oliver Lodge published detailed accounts of the apparatus used in his (Lodge's) public demonstrations of wireless telegraphy in 1894.

In 1914 Marconi was made a Senator in the Italian Senate and appointed Honorary Knight Grand Cross of the Royal Victorian Order in the UK. During World War I, Italy joined the Allied side of the conflict, and Marconi was placed in charge of the Italian military's radio service. He attained the rank of lieutenant in the Italian Army and of commander in the Italian Navy. In 1924, he was made a marquess by King Victor Emmanuel III.

Marconi joined the Italian Fascist party in 1923. In 1930, Italian dictator Benito Mussolini appointed him President of the Royal Academy of Italy, which made Marconi a member of the Fascist Grand

SSPD Chapter 1_Part 2_Photoelectric Effect

This chapter deals with the simple and elegent explanation given by Einstein in 1905 for Photo-electric Effect. For the experiments and the explanations provided by Einstein on this topic he was awarded Nobel Prize in Physics.

1.2. PHOTO-ELECTRIC EFFECT AND EINSTEIN EQUATION.

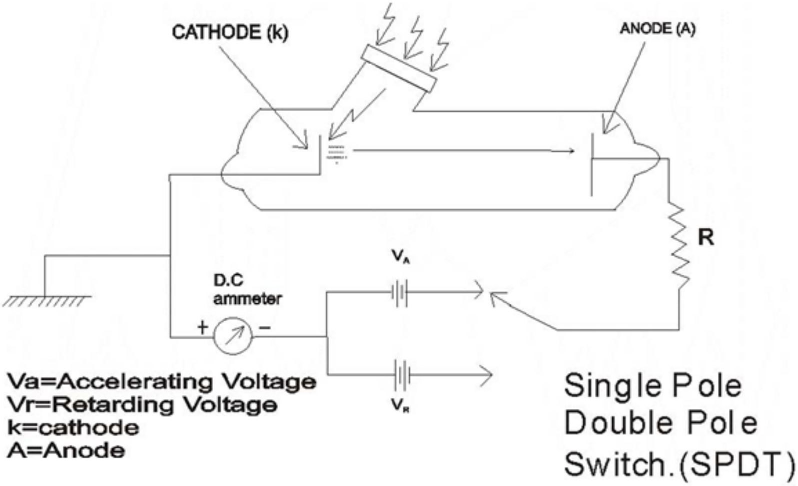


Figure 1.2.

V_A = Accelerating Voltage;

V_R = Retarding Voltage;

K = Cathode;

A = Anode;

W_F = Work-Function of the Cathode which is Thoriated Tunsten coated with Strontium Barium Oxide;

Photon Energy minus the Work-Function gives the kinetic energy of the photo-emitted electrons coming out of the cathode in the vacuum space surrounding the cathode;

$h\nu - W_F = (1/2)m_e v^2 = qV_R$; This Equation implies that if there is retarding voltage on the Anode then photo-emitted electron with a kinetic energy $(1/2)m_e v^2$ is totally expended in climbing up the Potential Hill created by the Retarding Voltage. Hence no photo current is detected by the Anode.

Fig.1.2. The Apparatus for verifying Photo Electric Effect and Einstein Equation.

Figure.1.2. describes the Experimental Set Up for verifying Photo Electric Effect and Einstein Equation. It consists of a Vacuum Tube which consists of a Cathode and an Anode. A Vacuum Tube is evacuated to an air pressure of 10^{-3}mm of Mercury which is referred to as 10^{-3} Torr. [1 Torr is the pressure which supports 1 mm of column of mercury. Torr is in the honour of the scientist Torrceli. Torrceli was the scientist who suggested that there was a near vacuum above the 76cm column of mercury supported by atmospheric pressure. When a 85cm tall glass tube , filled with mercury, is inverted over a trough full of mercury then only 76cm column is supported. The space above this column within the tube is empty except for mercury vapour which is negligible because of very high cohesive force between the molecules of Mercury. This empty space is referred to as Torcelli vacuum.]

An accelerating voltage or decelerating voltage can be applied to the Anode A Monochromatic beam of light of frequency ν is obliquely incident through a side tube on the Cathode for photo-ionic emission of electrons from the Cathode.

The Energy Band Diagram, which will be derived in Section 1.8., is given in Fig.1.3. As seen from the Diagram, the outermost unfilled Band is known as Conduction Band. At Zero Kelvin ,the Conduction Band is completely occupied till Fermi Energy Level E_F . As the temperature is raised , by Equipartition Law of Energy the Conduction Band electrons have (1/2)kT Average Thermal Energy associated with each degree of freedom and semi-free electrons in Conduction Band have three translational degree of freedoms. Hence the average thermal energy associated with the semi-free electrons in Conduction Band is (3/2)kT Joules. How the electrons are distributed above and below the Fermi-Level Energy at temperature T Kelvin will be clear when we discuss Fermi-Dirac Statistics.

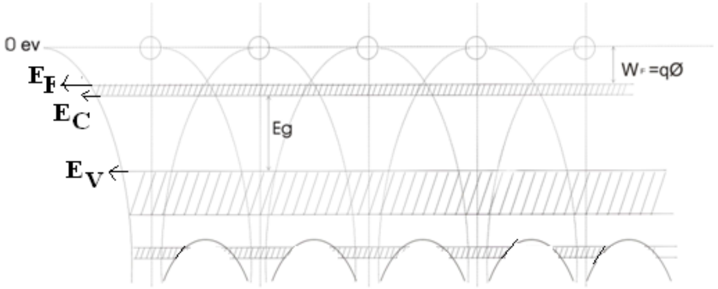


Figure 1.3. Energy Band diagram of Metal.

W_F=Work Function
$\varnothing_{BS}$ = Surface barrier potential

E F = Fermi level Energy. It lies within the band gap for metal.

E G = Energy Band Gap = Forbidden Zone.

E C = Lower Edge of the Conduction Band.

E V = Upper Edge of the Valence Band.

Fig 1.3. Energy Band Diagram of a Metal.

The electrons in the Conduction Band are not localized to their host atoms but belong to the whole crystalline lattice of the metal. Hence they are free to conduct throughout the crystalline lattice. But when they come to the surface boundary of the crystalline lattice they face a Surface Barrier Potential φ_B . This prevents the conduction electrons from escaping to the vacuum space outside the metallic lattice. Hence the conduction electrons are referred to as semi-free electrons.

But if the conduction electrons acquire sufficient energy to cross the Surface Barrier Potential then indeed the conduction electrons will be emitted from the surface of the Cathode.

The Surface Barrier Potential φ_B implies that conduction electrons must acquire energy equal to or greater than $q\varphi_B$ to escape into the vacuum. This quantity $q\varphi_B$ is known as the Work Function (W_F) of the Metal. Pure Tungsten has $W_F = 4$eV whereas Strontium Barium Oxide coated Thoriated Tungsten has $W_F = 1$ eV. The latter is used as the Cathode Material in all Vacuum Tube Devices.

Table 1.1. Work Function and Threshold Frequency.

Metal	Symbol	Work-Function(eV)	Threshold Frequency
Cesium	Cs	1.9	IR
Potassium	K	2.2	Red
Sodium	Na	2.3	Red
	Visible part of spectrum		
Tungsten	W	$\simeq 4$	UV
Aluminum	Al	4.08	
Lead	Pb	4.14	

Zinc	Zn	4.31
Copper	Cu	4.7
Silver	Ag	4.7
Platinum	Pt	6.4

If the conduction electrons become energetic enough to cross the Surface Barrier Potential by heating i.e. $(3/2)kT > W_F$ then *Thermo-ionic Emission* takes place.

If the incident Photons have energy packets $h\nu > W_F$ then *Photo-ionic Emission* takes place.

If the application of a very high field causes the surface barrier potential height to be lowered and surface barrier thickness to be narrowed then *High Field Emission* of conducting electrons occur by quantum mechanical tunneling. This is described in Section 1.7.3.

Presently we are concerned with Photo-ionic Emission.

Let the accelerating voltage be V_A ,

Let the incident photon have a frequency of ν ;

Let the Threshold frequency be ν_{Th} where $h\nu_{Th} = W_F$;

As long as incident photon frequency $\nu < \nu_{Th}$, no photo-excitation takes place and hence no photo-ionic emission takes place and no Anode Current is detected irrespective of what the Anode Voltage is. The increase in the intensity of light will also not result in any photo-excitation and no Anode Current will be ever detected as shown in Fig 1.4.

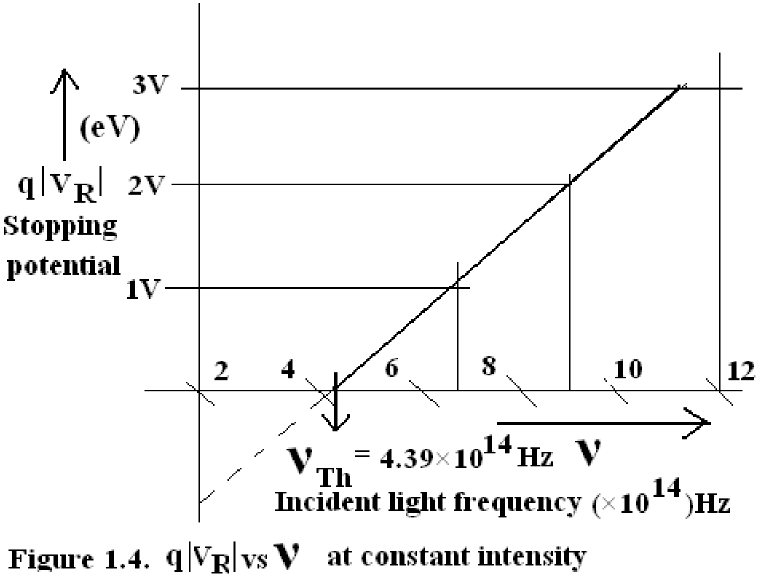


Figure 1.4. Stopping Potential, V R , versus incident frequency.

But if the incident frequency is greater than the Threshold Frequency then even the weakest intensity light at minimum Anode Voltage will cause some photo-ionic current and a weak Anode current will be detected.

As the intensity of incident monochromatic light is increased the strength of the Anode Current increases .

Now if a decelerating voltage is applied to the Anode, the photo current detected decreases until at a given negative Anode Voltage the Anode Current becomes zero. For constant frequency but varying Intensity, the stopping potential is the same. Meaning by that stopping potential depends on frequency and not on intensity as shown in Figure 1.5.

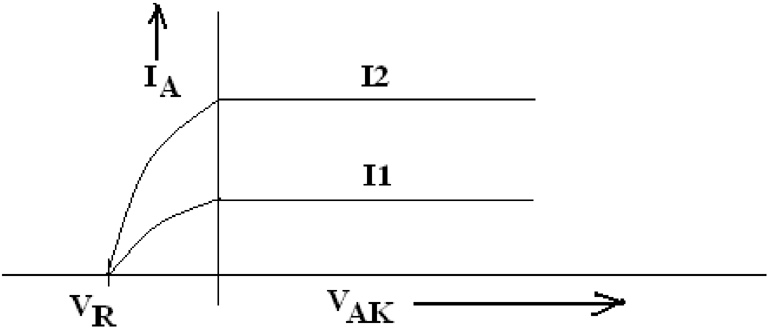
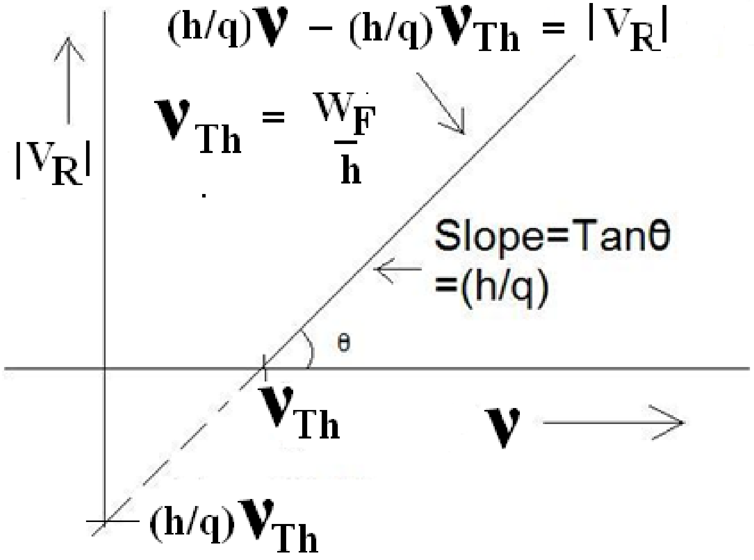


Figure 1.5. Photoionic Current I_A versus V_{AK} for the same monochromatic light.

On the basis of all the observations we arrive at the following graphs shown in Fig.(1.6), Fig.(1.7) and Fig(1.8).



Fig(1.6)

$h\nu - W_F = (1/2)m_e v2 = q\,|V_R|$where, $h\nu > W_F$,is the necessary condition for photoemission

$\nu - W_F/h = \nu - \nu_{Th} = (q/h)\,|V_R|$

or $(h/q)\nu - (h/q)\nu_{Th} = |V_R|$is a straight line equation like $y=mx+c$ where $|V_R|$ is the dependent variable y, ν is the independent variable x and (h/q) is the slope Tanθ where θ is the angle of inclination of the straight line.

Y axis intercept is $= -(h/q)\nu_{Th}$

Fig.1.6. Photo- electric effect graph with a constant intensity source but variable frequency.

From graph given in Figure 1.4 and 1.6 the Universal Constant h, Planck's Constant and Work Function of the metal of the cathode which is being illuminated by the monochromatic light can be determined. In 1930 Milikan, of Milikan drop experiment fame, carried out the experiment using Sodium as the cathode.

In his experiment he got the Plot as given in Figure 1.4.

According to that plot:

$\nu_{Th} = 4.39\times10^{14}$ Hz .

This corresponds to visible light of wavelength $\lambda_{Th} = 683$nm (red light).

The slope $= \Delta E/\Delta\nu = 4.1\times10^{-15} = h/e$.

Multiplying by the charge of electron $e = 1.6\times10^{-19}$ Coulomb ,

The value of Planck's Constant $= h = 6.6\times10^{-34}$ Joules.sec

[To be Continued in next module]

(SSPD_Chapter 1_Part 2 continued
This is continuation of Part 2_Photoelectric Effect.

Photo electric effect continued.

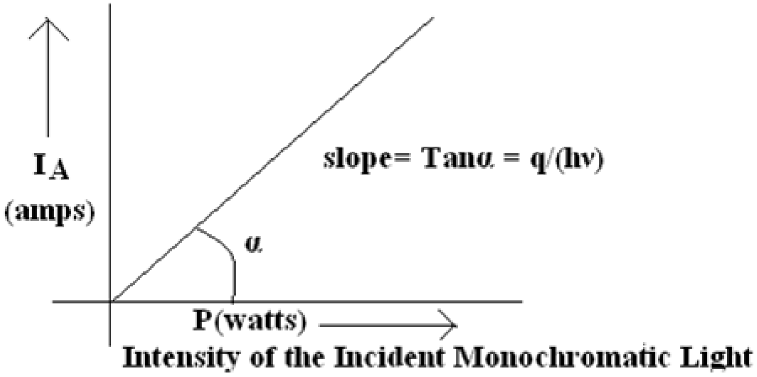


Figure 1.7. Photoionic Current(IA) vs Intensity of the incident monochromatic light.

P (Joules/second) / (hν) = number of photons incident .

If quantum efficiency is 100% then each photon will cause photoemission of a corresponding electron.

Hence P (Joules/second) / (hν) = number of photoemitted electrons/second.

Assuming that all photo-excited electrons are picked up by the Anode , this will constitute the Photo-ionic Current. Hence

I_A (Photo-ionic Current) = q.P (Joules/second) / (hν)

This graph is a straight line graph passing through the origin and with a slope Tanα=[q/((hν)] where α is the angle of inclination of the straight line graph.

Fig(1.7) Photoionic Current versus Intensity of Light for a monochromatic light source.

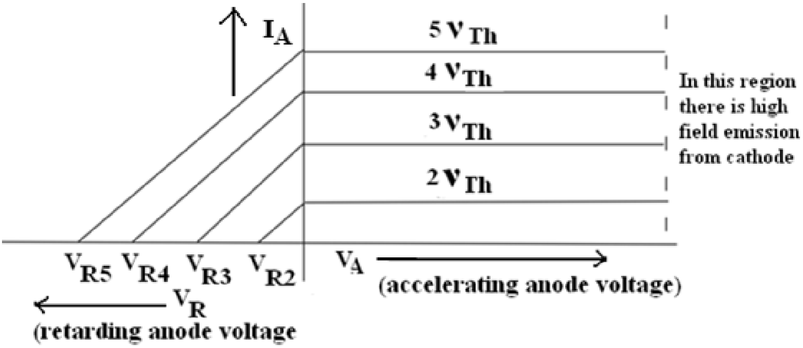


Figure 1.8. I_A . vs V_A for constant intensity but frequency of the incident light being increased in steps of ν_{Th}

Fig.(1.8) I A vs V A family of Output Curves for a constant Intensity but incident light frequency being increased in steps of ν Th .

Einstein was able to explain all the three graphs by assuming the quantum nature of light. Does this mean that Wave Nature of Light is wrong ? Absolutely not. We have studied Diffraction (aperture is comparable to wave length) and Interference pattern in Intermediate Physics and this can be explained only on the basis of Wave Nature of Light.

What this implies is that Electromagnetic Waves , which visible light is, manifests Wave Nature while interacting with Energy whereas it manifests Quantum Nature while interacting with Matter. Photo-Electric Effect and Compton Effect are two such examples.

The Photo-Electric also implies that every metal has its characteristics Work-Function and whenever the conduction electrons are energetic enough to overcome this Work-Function they will escape from the metal surface into vacuum.

Work-Function can be defined as follows:

It is the minimum energy required at zero Kelvin for a given metal for electron emission into vacuum.

A monochromatic light source emits photons of frequency ν at a constant rate of $P/(h\nu)$ photons per second where $P =$ Intensity of the monochromatic source in Joules per second. Each incident photon interacts with a single electron only and imparts the energy packet in totality to the electron with which it is interacting. If the incident photon's frequency is less than the Threshold Frequency then no intensity of light will cause photo-ionic emission. This is because increase in intensity means increase in the number of photons but energy packet transferred from the photon to the interacting electron remains insufficient to cross the surface potential barrier. Hence below threshold frequency no amount of incident light will cause any photo-ionic current.

When the frequency of the monochromatic source is increased to become equal to the threshold frequency then photo-ionic emission will just occur with photo-emitted electrons having zero kinetic energy.

When the incident frequency exceeds the threshold frequency then photo-emitted electrons have finite kinetic energy $(1/2)m_ev^2$ and photo-ionic current starts being detected as anode current. This kinetic energy can be measured by the application of retarding voltage at the anode. The application of positive Anode Voltage causes a potential downhill which accelerates the photo-emitted electrons. Hence positive Anode Voltage facilitates the collection of photo-ionic current as Anode Current. The application of negative Anode Voltage causes a potential uphill which retards the photo-emitted electrons and converts the kinetic energy of the photo-emitted electrons in potential energy and this prevents the collection of photo-ionic current as Anode Current and hence Anode Current decreases. As retarding voltage is increased the anode current decreases until at a retarding voltage V_R the anode current becomes zero. This gives the famous Einstein Equation:

$$h\nu - W_F = (1/2)m_e v^2 = q\,|V_R| \quad 1.5$$

where $h\nu > W_F$ is the necessary condition for photoemission

$$\nu - W_F/h = \nu - \nu_{Th} = (q/h)\,|V_R| \quad 1.6$$

$$\text{or } (h/q)\nu - (h/q)\nu_{Th} = |V_R| \quad 1.7$$

is a straight line equation like $y = mx + c$

where $|V_R|$ is the dependent variable y,

ν is the independent variable x

and (h/q) is the slope $\tan\theta$

where θ is the angle of inclination of the straight line.

The y intercept in Eq.(1.7) is $-(h/q)\nu_{Th} = -W_F/q$ = Surface Barrier Potential φ_B;

The x intercept in Eq(1.7) is ν_{Th} ;

Thus from the straight line graph $|V_R|$ vs ν (the incident photon frequency) , Fig(1.4), we can obtain Surface Barrier Potential φ_B and Planck's Constant h .

In Fig(1.7) the straight line graph I(Anode Current) vs P(Incident light Intensity) is obtained. Here the incident frequency, which is greater than the threshold frequency, is kept constant while the intensity of the incident light is increased.

At frequencies greater than threshold frequency, each incident photon causes corresponding photoemission of a single electron . Because of the one to one correspondence the Anode Current ,which is the collection of photo-ionic electrons per second, is directly proportional to the incident light intensity provided we have a constant positive Anode Voltage. This results in a straight line Graph passing through the origin.

Fig(1.8) is the family of curves comprising of the static output characteristics of Photo-Electric Device. The static output characteristic is I_A vs V_A for a given incident frequency and which is a multiple of threshold frequency and for a given intensity of light.

In the first quadrant of the Graph if we do not enter high field emission region then for a given incident frequency we have a constant Anode Current with respect to Anode Voltage provided the intensity of light is kept constant. Constant light intensity implies constant number of incident photons. If quantum efficiency is 100%, constant incident photons number implies constant photo-emitted electrons. Hence no matter what the anode accelerating voltage is, a constant number of photo-emitted electrons are collected. That is the Photo-Electric Current Device acts as constant current source as long as the incident light intensity is constant.

Here we may define Constant Current Sources and Constant Voltage Sources. A Constant Current Source delivers a constant current irrespective of the load voltage whereas a Constant Voltage Source will give a constant terminal voltage irrespective of the load current.

The I-V curve of a Constant Current Source will be a horizontal line as it is in the case of Photo-Electric Current Device whereas the I-V curve of a Constant Voltage Source will be a vertical line.

For this simple but elegant explanation of Photo Electric Effect , Albert Einstein was awarded Nobel Prize in 1921.

These photons are virtual particles but they exhibit all the properties of material particles. They have translational momentum and hence exert force and pressure. Had the pressure of sun-light not been considered, Viking would have missed Mars by about 15,000 km.

Insolation in the upper most part of Atmosphere= 1400W/m2;

Force (F) = rate of change of translational

momentum (dp/dt) 1.8

Force.displacement = F.Δx

= ΔU (incremental work done on incremental displacement)

F.Δx = (dp/dt). Δx = Δp.(dx/dt) = Δp.v 1.9

In case of light in free space : ΔU = Δp.c;

Therefore Δp = ΔU/c;

Therefore Δ**p/Δt** = **ΔU/Δt /c** = **(1/c)(dU/dt)** = **Force** = **F (Newtons) 1.10**

dU/dt = rate of energy incidence;

dU/dt/m2 = Energy falling per unit time per unit area = Insolation;

Eq.(1.10) gives:

F/m2 (Pressure)= (1/c)(dU/dt)/m2 = I (intensity)/c ;

Given I(intensity) =

Insolation of sun-light in the upper part of the atmosphere = 1400W/m2

Resulting pressure= I (intensity)/c

$= 1400\,W/m^2 \div$ (velocity of light in vacuum)

$= 4.7 \times 10^{-6}\ N/m^2$;

$1\,mW/m^2$ is equivalent to $3.3\,pN/m^2$;

Total force exerted on Earth by Sunlight $= (F/m^2) \times (\pi R_{EARTH}^2)$

$= 4.7 \times 10^{-6}\ N/m^2 \times \pi \times (6.37 \times 10^6)^2 m^2$

$= 6 \times 10^8\ N = 6 \times 10^8\ N/(10^4 N/\text{metric Ton})$

= 60,000 metric tons.

Theory of Dualism as proposed by the French Prince de Broglie

The French Prince had proposed the theory of Dualism. If light can act as a particle then why electron cannot behave like wave. Davisson and Germer went on to test this hypothesis by firing an electron beam on a single crystal of Nickel. Electron got diffracted in exactly the same manner as X-Ray does in Crystallography experiment.

1.3. PROPOSITION OF THE THEORY OF DUALISM BY THE FRENCH SCIENTIST DE BROGLIE.

If Energy Waves could manifest particle like properties then why material particles could also not behave like matter waves? This was the question posed by the French Scientist de Broglie **[Appendix XIV]** .

He proposed that the wavelength of the matter wave is the following:

$p = h/\lambda$ 1.11

where $p =$ translational momentum of the material particle,

$h =$ Planck’s Constant,

$\lambda =$ Wavelength of the matter wave under consideration;

If an electron is accelerated through (N) kVolt then the Kinetic Energy acquired is:

$qV = N \times 10^3 eV = N \times 10^3 \times q$ Joules

$= 1.60218 \times 10^{-16} \times N$Joules $= (1/2)m_e v^2$

Therefore linear momemtum acquired is :

$p = m_e \times v = (\sqrt{N})1.7085 \times 10^{-23}$ Kg.m.sec^{-1}

Therefore: $\lambda = h/p = [0.388/(\sqrt{N})]$ Angstrom.

The wavelength of electron matter particles with 10 keV is of the order of 1/10 Angstrom

(10^{-10}m).

For this radical theoretical conceptualization de Broglie was awarded the Nobel Prize in 1929. This became the basis for the invention of Electron Microscope (Section 1.12).

The optical microscopes have a maximum magnification of 2,000 times whereas Electron Microscopes have magnifications of 2,000,000 times

The resolution capacity and hence the magnification power directly depends upon the wavelength of the interacting wave. Visible light wavelength ranges from 4000Å(Blue) to 7000 Å(Red) whereas Electron has a wavelength of 1 Å. Therefore Electron

Microscope has 4000 to 7000 times more resolution as compared to the Optical Microscope.

Proton’s matter wave has still shorter wavelength hence it can resolve still finer features.

In Geneva CERN lab, in Fermi Lab of Massachusetts and at SLAC in Stanford Proton particles are being accelerated to higher and higher velocities to smash the nuclei of different elements. This will help reveal the inner structure of the nuclei and also recreate the events which followed in the first fractions of a second right after the Big Bang Explosion of Creation.

Till date Tera Electron-Volt(10^{12}eV) energy has been achieved in these particle accelerators and three basic families of elementary particles have been discovered. They are tabulated in Table(1.1). In CERN, Geneva, 7Tev energy has been reached by Large Hadron Collidor (LHC).

The most recent announcement by “ High Energy & Nuclear Physics “ group at Brookhaven National Laboratory (Times of India, Patna Edition, Wednesday, 20th April,2005) states that the collision of two Gold Ions at speed of light recreates 10^{12} Kelvin (trillion Kelvin temperature) a condition present in first microsecond after Big Bang. This Tera electron-volt energy collision experiment released the quarks from their confinement and showed them the existence of free quarks. The free quarks and gluons existed as plasma and not in gaseous state. Thus as we go higher up in high physics realm we are able to look at the finer features of matter as well as its very elementary form.

Table 1.1. The Three Basic families of elementary particles.

Leptons	Anti-leptons	Quarks	Anti-Quarks	Quarks	Anti-Quarks	Neutrinos	Anti-Neutrinos
Electrons	Electrons*	Strange	Strange*	Charm	Charm*	Electron-neutrino	Electron-neutrino*
Tau-leptons	Tau-leptons*	Down	Down*	Up	Up*	Tau-neutrino	Tau-neutrino*
Muons	Muons*	Bottom	Bottom*	Top	Top*	Muon-Neutrino	Muon-Neutrino*

1.3.1. CONDITIONS FOR EXHIBITING WAVE-PARTICLE DUALITY-DEGENERACY TEMPERATURE.

We know from diffraction experiment and double slit Young’s Interference Experiment **[Appendix**

XV] that when the limit of Geometrical Optics is reached, ray optics is replaced by wave optics. That is when the wavelength of light(λ) is much smaller than the aperture dimension ‘a ‘ light is not diffracted and Ray Optics is applicable. When the aperture dimension is comparable to the wavelength then diffraction starts taking place.

What this means is that light shows its particle nature when aperture is very wide but light exhibits wave nature when aperture is of comparable dimension.

In a similar manner when particle’s de Broglie wavelength is comparable to the spacing between the scatterers then wave nature of the particle is exhibited.

In quantitative terms: $(mv_2)/2 = (3/2)kT$

Therefore $mv = \sqrt{}(3mkT) = p$ (linear momentum);

Therefore the condition for wave particle duality is:

$\lambda = h/\sqrt{}(3mkT) \approx a$ (spacing between the scatterers);

At **T ≤ T0 = h2/(3mka2) 1.12**

= **Degeneracy Temperature where we have wave-particle duality**;

For solids and liquids a = 2 to 3 Angstrom

Therefore an electron which has mass of the order of 10^{-30}Kg will manifest wave nature below degeneracy temperature $T_0 = 265\times 10^3$K. That is an electron in a solid/liquid will exhibit wave nature for all temperature at which matter can exist in solid/liquid phase. Therefore Electron Microscope operates at Room Temperature.

If electron is replaced by an hydrogen atom and scatterer spacing is of the same order then T_0 is 50Kelvin. Therefore wave nature of hydrogen atom will show up below 50K . Super fluids of helium are a result of wave particle duality. Therefore we have to achieve quantum liquids where atoms exhibit wave nature and this can be achieved only at liquid helium temperature or lower.

1.4. ELECTRON DIFFRACTION BY CRYSTALS (THE FINAL VINDICATION OF MATTER WAVES)

In 1927 an accident in Bell Telephone Lab (founded by Alexander Graham Bell **[Appendix XVI]**) went a long way in establishing the revolutionary concept of matter waves.

In 1909 Earnest Rutherford **[Appendix XVII]** was carrying a systematic research on the atomic structure. In the process he bombarded a thin gold foil with Alpha Rays. He recorded the deflection of the Alpha particles. While studying the deflection pattern he concluded that most of the mass of the Atom was concentrated in the nucleus of the Atom and electrons were orbiting the nucleus in much the same way as planets were orbiting the Sun. Thus he proposed the Planetary Model of Atom.

In 1927 like Rutherford, Josef Davisson and Lester Hailburt Germer(in USA) **[Appendix XVIII]** were bombarding Nickel foil with electron beam. Initially Nickel was a poly crystal and there was nothing noteworthy in the scattering pattern and the scattering pattern was no way correlated to the accelerating voltage in the electron gun. But then an accident occurred and air leaked into the apparatus. Nickel got oxidized and to reduce it to elemental form it was baked in an oven. In the course of the heating cycle , Nickel had transformed into a large single crystal structure and this periodic structure diffracted the electrons in much the same way as X-Ray had been diffracted in X-Ray Crystallography developed by Bragg **[Appendix XIX]** . The apparatus , the scattering pattern and its dependence on voltage is shown in Figure(1.7) [*This experiment is described in Modern Physics by Richtmeyer, Kennard & Cooper ,1969, McGraw Hill*]

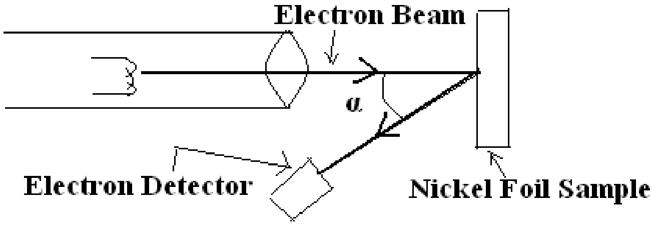
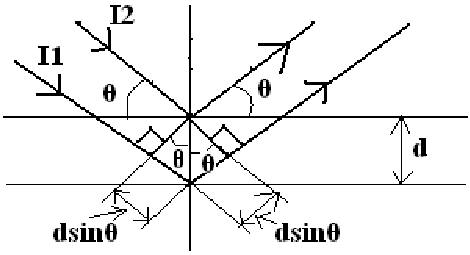


Fig.1.9.A. Experimental arrangement of Davisson Germer Experiment.



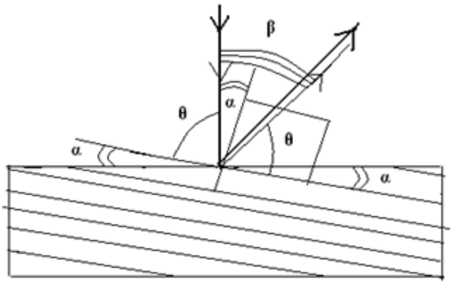
Bragg's condition for constructive interference

$n\lambda = 2d\sin\theta$

path difference between the two rays is $2d\sin\theta$

When path difference is integral multiple of wave length we have constructive interference

Figure 1.9.B



Diffraction of electron waves from the Bragg's Planes in the crystal of Nickel. Electron was accelerated by 54V hence it had a Kinetic energy of 54eV which corresponds to 1.66Angstrom. The crystal lattice parameter = d = 0.91Angstrom.

As seen from the diagram a normally incident beam of electrons will give a diffration maxima at 50degree to the normal as shown in the diagram.

$\beta = 50^\circ$ $\theta = 65^\circ$ $\alpha = 25^\circ$

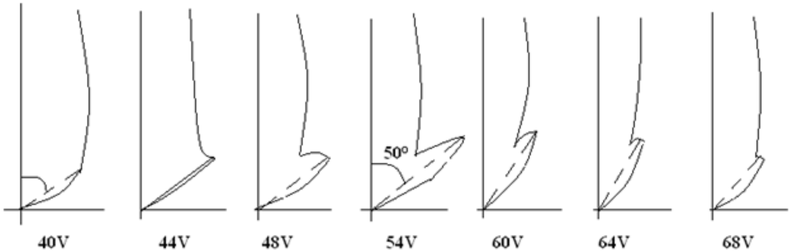
Diffraction maxima at 50° to the normal for a normally incident beam of electrons is exactly in accordance with Bragg Equation for constructive interference namely:

$n\lambda = 2d\sin\theta \longrightarrow$ this gives θ as 65°

This is the angle which incident beam makes with the parallel crystalline planes.

The reflected beam makes the same angle of 65 degree with respect to crystalline planes.

FIGURE. 1.9.C.



The voltages indicated along the X-axis are the accelerating voltages of the incident electron beam.

The electron beam is detected at different azimuthial angles. For 54V accelerating voltage,scattering maxima occurs at 50 degree to the normal to the surface of the crystal.

Figure 1.9.D

Figure 1.9. The Davisson-Germer Experimental Appratus and the resulting scattering pattern and its dependence on voltage.

From the diffraction pattern, the wavelength of the incident waves could be determined if the lattice constant was known. Also detailed crystalline structure could be determined.

Since lattice constant , wavelength of electrons and wavelength of X-Ray are of the same order, of the order of Angstrom, hence Crystallography can be carried out as easily with electrons as they are carried out with X-Ray.

For X-Ray crystallography, Lau and Bragg Father and Son Group received Nobel Prizes in 1914 and 1915 respectively.

For confirming the existence of Matter Waves by electron crystallography, Davisson(USA) , Germer(USA) and G.P.Thomson (England) **[Appendix XX]** received the Nobel Prize in 1937. G. P. Thomson, son of J. J. Thomson the discoverer of electron, independently verified the wave-particle duality of electron.

In 1931 Stern and Collaborators carried out the matter wave diffraction using slow He atoms targeted on LiF crystals. Again Lau diffraction pattern was obtained.

In 1988, Keith et al (Physical Review Letter, Vol. 61, pp. 1580) repeated the experiment with slow Na atoms(11e,11p & 12n) targeted on Gold Membrane with two well separated slits. This gave the interference pattern as obtained in Young’s double slit interference experiment.

In 90’s neutron interferometer has been achieved. The neutron beam is passed through well separated slits and after the passage through the slit it is allowed to fall on a screen where incident neutrons are counted at different points of incidence in transverse direction. We obtain alternating dark and bright fringe pattern as obtained in Young's double slit experiment on light. Here dark implies low count and bright implies high count. If one of the slits is closed we get uniform pattern. The dark and bright fringes disappear.

In 1666 Issac Newton **[Appendix XXI]** had proposed the corpuscular nature of light. Only in 1801 when Thomas Young demonstrated the phenomena of Interference in monochromatic light that wave nature of light was accepted and established.

In 1927 the diffraction pattern of electrons finally established the wave nature of matter with definiteness.

SSPD Chapter 1_Part 3_Theory of Dualism_Matter wave continued

SSPD_Chapter 1_Part3 had described the hypothesis and corroboration of Matter Waves. Here we give the form of expression of matter wave and its physical interpretation.

Chapter 1_Part3_ de Broglie’s matter waves continued.

De Broglie’s hypothesis was verified by Davisson & Germer conclusively. Here a material particle is behaving like a light wave - what does this matter wave mean ? What does this matter wave represent ? The matter wave represents Probability Amplitude Ψ(r,t) and |Ψ|2 represents probability density and |Ψ|2.dV represents the probability of finding the matter particle in elemental volume space dV.Matter wave or electron wave is not smeared matter. It is still particulate matter but the probability of location is smeared, is wave like. What this means is that there is no certainty of finding the particle in one location.In macroscopic world or in classical world, a body at rest has probability of finding it in one given location as UNITY and elsewhere the probability of locating it is ZERO. Unity means certainty.Zero means total absence.But in microscopic world or in quantum mechanical world this is not the case.The probability is spread all over the space within a certain bound.

$$\Psi(r,t) = \Psi_0 \exp\left([j(\omega t - kr)]\right)$$

- this is the standard format in which a travelling harmonic wave is represented.

$$\Psi_0 \exp([j(\omega t - kr)])\ \mathit{reprsents\ a\ forward\ travelling\ wave.}$$

$$\Psi_0 \exp([j(\omega t + kr)])\ \mathit{reprsents\ a\ backward\ travelling\ wave.}$$

$$\iiint_{S1}^{S2} \Psi_0 \exp([j(\omega t - kr)])\, dV = k$$

Here k is less than Unity and gives the probability of finding the particle in a volume bounded by S1 and S2.

If a particle is trapped in volume V1 then the probability of finding the particle in volume V1 is Unity.

$$\iiint\limits_{V1} \Psi_0 \exp([j(\omega t - kr)])\, dV = 1$$

If a particle is outside the volume V1 then the probability of finding the particle in volume V1 is zero.

$$\iiint_{V1} \Psi_0 \exp([j(\omega t - kr)])\, dV = 0$$

In fact in Davisson and Germer experiment the dark and bright ring represent the number of events of detection. Dark represents very low number of events being detected and bright ring represents a large number of events being detected. This implies high probability of finding the particle in the bright region and low probability of finding the particle in dark region.

SSPD Chapter 1_Part 4_Heisenberg Uncertainty Principle

Chapter 1, Part 4 describes the Uncertainty Principle as proposed by Heisenberg and its application to prediction of Hawking Radiation and Angular Momentum of Fermions and Bosons.

1.5. WERNER HEISENBERG AND THE PRINCIPLE OF UNCERTAINTY.

Uncertainty Principle states that no two canonical conjugate variables can be fully determined with 100% accuracy simultaneously. The two canonical variables are position (x,y,z) and momentum(p) or energy(E) and time (t). What this means is that in process of making measurement of one canonical variable the other variable gets altered and visa versa hence the two cannot be simultaneously determined.

If uncertainty in spatial measurement is Δx,

And uncertainty in momentum measurement is Δp_x, then according to Heisenberg's Uncertainty Principle **[Appendix XXII]** :

$\Delta x.\ \Delta p_x \geq (h/(2\pi))$ or $\hbar$;

Similarly $\Delta y.\ \Delta p_y \geq (h/(2\pi))$;

Similarly $\mathbf{\Delta z.\ \Delta pz \geq (h/(2\pi))}$**; 1.9.**

In 1930 Einstein extended this Principle to Energy and Time and he stated that if the uncertainty in Energy measurement is ΔE,

And the uncertainty in Time measurement is Δt,

Then $\Delta t. \Delta E \geq (h/(2\pi))$. **1.10**

Uncertainty Principle laid the foundations of Fermi-Dirac Statistics and Solid State Physics.

Uncertainty principle tells us that energy involved in studying at Planck Length , which is $(Gh/c^3)^{1/2} = 10^{-35}$ m, is 10^{18}GeV. This is the kind of energy involved in studying Quantum Gravity. At 10^{15}GeV the strong , weak and electromagnetic unify in one force as will be seen in Section 1.13.

In 1974, Stephen Hawking **[Appendix XXIII]** conjectured on the basis of Uncertainty Principle that Black Holes give off radiation. He concluded that Black Hole looks like a hot body emitting particles at a temperature inversely proportional to the mass of the black hole. According to Special Theory of Relativity combined with Quantum Field Theory, in a static gravitational field virtual particles (particle and antiparticle) are being created and almost simultaneously annihilated. Heisenberg Uncertainty permits this kind of creation-annihilation without any expenditure of energy.

Suppose this kind of thing is happening near the event horizon of Black Hole. Event Horizon is the ‘one way membrane’ through which nothing can escape. Now near the event horizon it is always possible that antiparticle gets trapped and particle escapes. There is negative energy associated with antiparticle and positive energy with particle. In effect, the radiation is taking place at the cost of the mass of the Black Hole. If this Radiation continues for a long time the black hole will wither away. This radiation has been called Hawking Radiation and is yet to be detected..

Just as we have Uncertainty between E and t or p and r, so we have uncertainty between angular momentum and angle θ.

In p-r, p can have any value from 0 to ∞ and r can also have any value from 0 to ∞. But θ can vary only from 0 to π hence spin angular momentum can vary as multiples of ħ.

As we know electron spin angular momentum can vary between (-1/2 ħ) and (+1/2 ħ) and W± and Z_0 can have spin angular momentum values as – ħ, 0 and ħ. W± and Z_0 are intermediate vector bosons which mediate weak forces. Since W± and Z_0 are bosons hence they have spin angular momentum which is an integral multiple of ħ .

Story of Atom_Plum-Pudding Model of Atom as proposed by J.J.Thomson.
J.J.Thomson by the various kind of studies of Cathode Ray was able to convincingly prove that it indeed was electron which had negative charge and negligible mass compared to that of the nucleus. ,

1.6. THE STORY OF ATOM AND THE BOHR MODEL OF ATOM.

The Story of Atom is really the search for the Philosophy of Life. What is the purpose of Human Existence or if indeed there is a purpose in our existence. The Greeks first raised the question of the ultimate unit of matter and presented a possible model of the same.

Around 450BC Liusipus of Miletius and Democretus of Abdera proposed the following Model of Atom:

Matter has a particulate nature and if we go on dividing and further subdividing it we come to an indivisible portion of Matter and this indivisible portion is known as Atom.

Plato and Aristotle, the great Philosophers of the Greek Era, did not subscribe to this view.

With the passage of time the question about Atom kept burning and Human Mind remained ever inquisitive about the same.

The inquisitive enquiry led to a series of experiments performed by Dalton in 1801 in England. On the basis of these landmark experiments Dalton proposed the following:

Every element is made of identical Atoms;

Different elements are constituted of different Atoms;

The Atoms of different elements chemically combine to form a third kind substance called Compound which is completely different chemically as well as physically from the constituent elements;

In 1869, Damitri Ivanovich Mendeliv of Russia and Julius Lothar Meyer of Germany proposed the Periodic Table.

Periodic Table strengthened the conceptualization of Atom but what decides the chemical properties of an element remained an enigma.

The answer to this enigma of Chemistry was linked with the research on Electron and hence it became necessary to tell the Story of Atom here.

Till the end of Nineteenth Century, Atom was regarded as indivisible.

1.6.1. THE PLUM PUDDING MODEL OF ATOM AS PROPOSED BY J.J.THOMSON.

In 1854 Heinrich Grislier of Germany invented a high vacuum glass tube. The Scientists of the time inserted metal electrodes as Cathode and Anode. A very high Electric field was applied between the two electrodes and an electric discharge was achieved. During the electric discharge, the cathode was found to be surrounded by greenish glow. Goldstein of Germany regarded this Cathode Ray to be like electromagnetic wave. But other Scientists regarded this Cathode Ray to be particulate in nature. Particulate nature was proposed because electrons underwent classical electrostatic as well as magnetostatic deflection.

Finally Joseph John Thomson **[Appendix XX]** carried out a series of experiments on the Cathode Ray produced in a vacuum tube during an electric discharge and conclusively proved that Cathode Ray is composed of Electrons ,the constituent particle of Atom.

He studied the effect of Electric field as well as that of Magnetic field. Both caused deflection in the Cathode Ray. These experiments gave the charge of Electron to be negative and these experiments determined the ratio q/m_e to be $(1.6\times10^{-19}/9.1\times10^{-31})$. Electron was found to be thousandth part of Hydrogen Atom by mass.

Because of this light weight Electron is strongly deflected in a weak electric field.

J. J. Thomson studied both electric discharge vacuum tube as well as Photo Electric Vacuum Tube. He conclusively proved that in both the vacuum tubes same electron particles of negative charge are emitted.

This proved with definiteness that Atoms are not indivisible but can be further split into Electrons. Hence a negative charged electron is definitely a constituent particle of Atom. Since Atoms are electrically neutral unless ionized forcibly , it is logically correct to assume that there are positively charged matter too in the Atom. On the basis of these reasoning, J. J. Thomson proposed his Plum-Pudding Model.

According to this Model, Atom has a sea of positive charge and in this sea of positive charge the negatively charged Electrons are interspersed just as plums are interspersed in a pudding.

A high Electric Field or optical light above the threshold frequency will cause identical electrons to be emitted in the two cases.

For the complete exposition of the mysteries of Electron, J. J. Thomson was awarded the Nobel Prize in 1906.

SSPD Chapter 1_Part 5_Story of Atom-Solar System Model of Atom as proposed by Rutherford.

SSPD_Chapter 1_Part 5 describes the historic experiment of "Bombarding the Gold Foil with Alpha particles". This experiment established that most of the atom is empty and all the positive part is concentrated in the nucleus which contains practically all the mass also and that negatively charged electron are orbiting the nucleus just as Earth is orbiting the Sun.

1.6.2. SOLAR SYSTEM MODEL OF ATOM AS PROPOSED BY ERNEST RUTHERFORD.

From 1906 to 1908 by particle collision and particle scattering, Earnest Rutherford proved that nearly 99.9% of the mass is concentrated in the nucleus of the Atom and that the diameter of the nucleus is 4 orders of magnitude smaller than the diameter of the Atom. The diameter of the Atom is of the order of Angstrom(10^{-10}m) whereas the nucleus is of the order of 10 Fermi Unit where 1 Fermi Unitis 10^{-15}m. Since Atoms are electrically neutral hence nucleus contains as many positive charges as are the electrons orbiting the nucleus.

From this Model, most of the Atom is vacuum. Just as the Planets are orbiting the Sun, in an analogous manner electrons are orbiting the nucleus. Furthermore the nucleus contains as many positively charged protons as the number of electrons are orbiting the nucleus.

In the particle scattering experiment a Gold or Platinum thin foil was bombarded with Alpha particles. Most of the particles passed through the foil without experiencing any deflection. It appeared as if the interatomic and most of the intra-atomic space in the foil was empty space.

Though most of the Alpha particles passed through unhindered there was a minority group of Alpha particles which did experience small deflection and there was a rare few which were completely reversed in their track. The reversal in their track was tantamount to a deflection of 180° as shown in Fig(10).

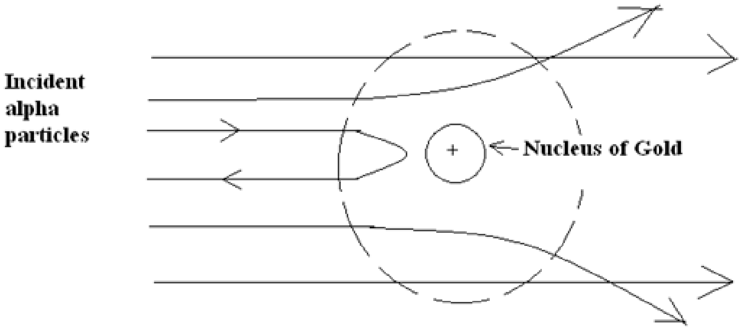


Figure 1.10. The scattering phenomena of Alpha Particles by a thin Gold Foil.

Figure 1.10. The scattering phenomena of Alpha Particles by a thin Gold Foil.

Since the electrons are 1838th part of proton by mass therefore the orbiting electrons have no influence on incident Alpha particles which ,as will be seen ,are the nuclei of Helium Atoms. But if the incident Alpha particle experiences a direct collision with the nucleus, the former is reversed in its track. This clearly proves that Alpha particles are of comparable mass as that of the nucleus.

Here it will be appropriate to mention that Alpha , Beta and Gamma Rays accompany Radioactivity. Gamma Rays are very short energetic Electromagnetic Waves of wavelength (1/100)Å, Beta Rays are electron beam whereas Alpha Rays are the nuclei of He atoms with two protons and two neutrons.

On the basis of α particles scattering by gold foil, Rutherford proposed that most of the atom is empty space with 99.9% of the total mass concentrated at the center called nucleus and electrons are orbiting around the nucleus. The negative charge of the orbiting electrons is exactly balanced by the positive charge of the nucleus.

In 1908 Earnest Rutherford was awarded the Nobel Prize in Physics for unraveling the planetary model of the Atom.

SSPD_Chapter1_Part6_Neil Bohr's Quantum Theory of Atoms.

SSPD_Chapter1_Part 6 gives the Quantum Theory of Atoms as proposed by Neil Bohr.

SSPD_Chapter 1_Part 6_NEIL BOHR’s QUANTUM THEORY OF ATOMS- STATIONARY AND RADIATION LESS ORBITS.

According to Classical Physics any accelerated charge will radiate electromagnetic waves called Synchrotron Radiation .

Electrons in circular orbits around the nucleus experience a centripetal force due to electrostatic attraction $((q.q)/(4\pi\varepsilon_o r^2))$ and a reaction force called centrifugal force equal to mv^2/r. An exact balance between the two keeps it in a stable circular orbit. If energy is being lost through radiation , due to the fact that a particle in a circular orbit is being continuously accelerated towards the centre, then the orbiting electron will spiral inward until it collapses into the nucleus. But no such thing happens in reality in an Atom though we do witness the phenomena of Synchrotron Radiation in a Betatron.

As shown in Fig.(1.11) , Betatron consists of a powerful magnetic field along z axis perpendicular to the plane of paper. The electron is accelerated along an expanding spiral path in the plane of the paper around the longitudinal axis of the magnetic field. An electric field is applied parallel to the plane of paper by two oppositely charged D shaped electrodes which accelerate the electron every time it passes from one electrode region to the other region. The electric field is alternating field with period of alteration equal to the time taken to cover one circular path. This keeps accelerating electron into expanding spiral path.

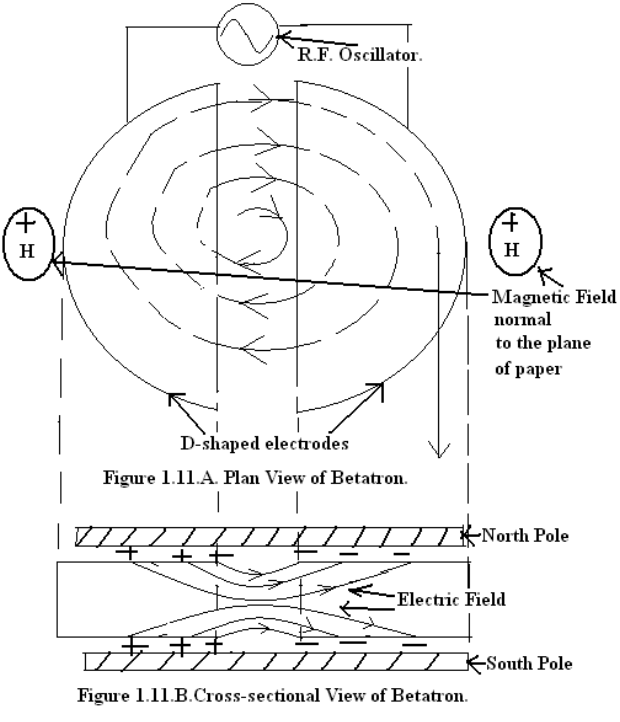

R.F. Oscillator.
+
H
+
H
Magnetic Field normal to the plane of paper
D-shaped electrodes
Figure 1.11.A. Plan View of Betatron.
North Pole
Electric Field
South Pole
Figure 1.11.B.Cross-sectional View of Betatron.

Fig.1.11. The schematic of a Betatron

The outward spiraling electron radiates a continuous spectrum of soft X-ray electromagnetic waves. This finds several commercial applications in the field of medicine and industries.

But what is true at macro scale is not true at the atomic scale. Why?

In 1913 Niels Henrik David Bohr(1885-1962) **[Appendix XXIV]** proposed a new set of postulates to circumvent this difficulty:

1. Electrons are permitted only in those orbits which are radiation less and support standing waves of electron probability amplitude $\Psi(r,t)$;

(ii) Electron Orbital Angular Momentum = integral multiple of $(h/(2\pi))$;

i.e. $I\omega = n\hbar$ 1.15

where I = moment of inertia around the spin axis = mr^2 ;

and ω = orbital angular velocity = v/r;

1. A photon is emitted when electron jumps from excited state to ground state. A photon is absorbed when electron jumps from ground to a excited state. In both the cases, that is in the case of radiation or absorption, the photon's energy packet $h\nu = \Delta E = E_2$ (excited state) – E_1(ground state);

Neil Bohr himself did not prove these postulates. Heisenberg used Quantum Mechanics to prove these postulates and Schrödinger used Matrix Mechanics to prove the same. Heisenberg was awarded Nobel Prize in 1932 for his contribution to Quantum Mechanics.

According to Matter Wave Theory, an electron in free space has a continuum of energy. It behaves as a traveling wave and its wave number and orientation can be continuously varied. This electron can occupy any energy level from zero to infinity.

Total Energy = E = Kinetic Energy + Potential Energy;

Therefore $E = p^2/(2m_e) - qV(r)$ 1.16

Where p = linear momentum of the electron;

m_e = mass of the electron;

q = the charge of an electron;

V(r) = potential at position r .

Wave number or wave propagation vector $k = 2\pi/\lambda = (2\pi/h)p = p/\hbar$

Or $p = k\hbar$ 1.17

An electron in free space is an unbounded electron and can assume a continuum of energy values as p can assume a continuum of linear momentum values. p can assume a continuum of linear momentum values because in free space there are no boundary conditions to restrict the linear momentum values.

From Eq.(1.16) E vs p or E vs k is a parabolic curve as shown in Fig.(1.12).

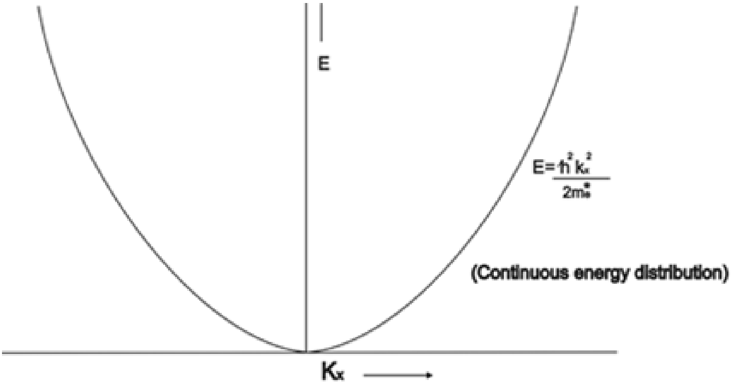


Figure 12. E-k diagram of electron in free space or in un-bounded space.

Figure.1.12. E vs k curve for an electron in free space.

$$E=\frac{p^2}{2m_e^*}=\frac{k^2\hbar^2}{2m_e^*}$$

Where m_e^* = effective mass of the electron.

When an electron is restricted to an infinite potential well [Figure 1.13.C] then the matter wave as a traveling wave is converted to a standing wave. In a potential well only those standing waves are permitted which satisfy the condition of nodes at the potential well boundaries just as a vibrating spring can vibrate only at those notes where the standing waves have nodes at the two ends of the string which are tied to the side walls. That is the wavelength of the standing wave must satisfy the condition:

$W = (n \,.\, \tfrac{1}{2})\lambda$. 1.18

Where $n = 0, 1, 2, 3$; and W is the width of the potential well.

Fig(1.13.a and b and c) give the permitted standing wave pattern for a vibrating string tied on the two ends as well as for an electron in One-Dimensional Potential Well with infinite potential barrier at the two ends and the figure of an infinite potential well respectively..

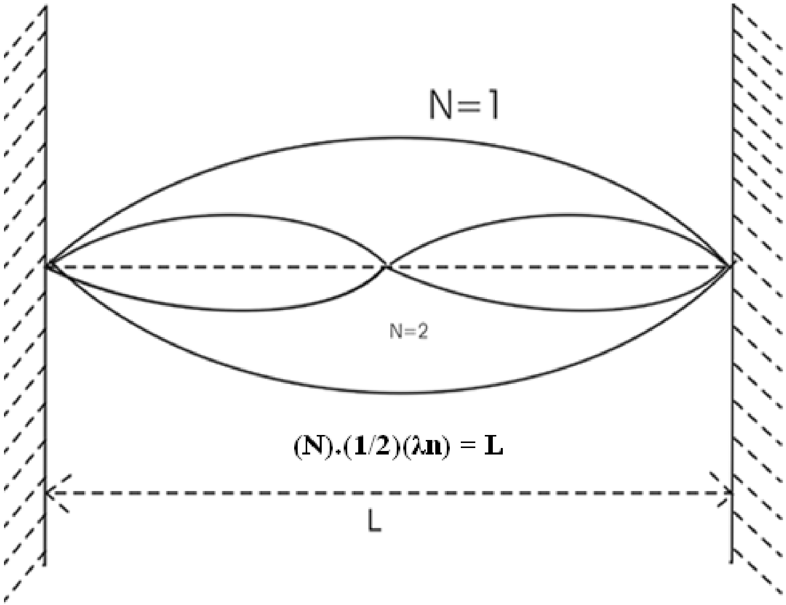


Figure 13.A. The Standing Wave Pattern for a vibrating string tied on the two ends;

Fig(1.13.A) The Standing Wave Pattern for a vibrating string tied on the two ends;

For a standing wave on a vibrating string tied on two ends, the following condition has to be satisfied:

$$(N).(1/2)(\lambda\ n) = L \quad 1.19$$

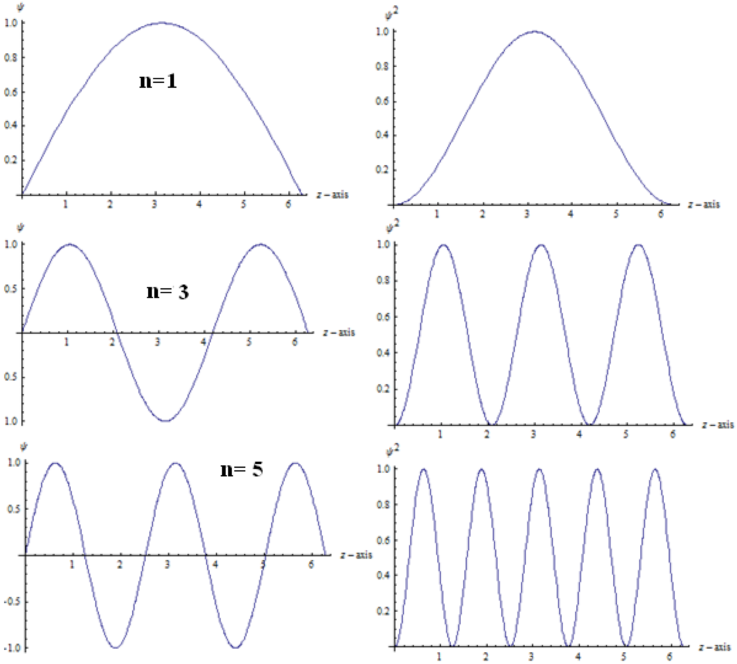


Fig.(1.13.B) The Standing Wave Pattern for an electron in 1-D infinte Potential Well in 1st,3rd,5th quantum states.

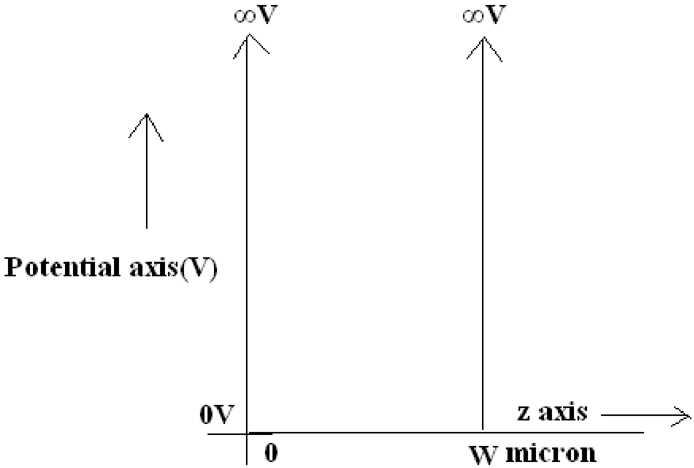


Figure 1.13.c. An infinite Potential Well of width W micron

Figure 1.13.C. An infinite potential well of width W microns

For standing wave in an infinite potential well, the following condition has to be satisfied:

$(n.1/2)(\lambda n) = W$ 1.20

The positively charged nucleus of an Atom creates 3-D Potential Well of spherical symmetry. In this finite space, electrons can exist only as Standing Waves with the boundary condition:

$2.\pi.r = n.(h/p n) = n.\lambda n$ 1.21

where n = 1, 2, 3, 4,……….

Eq.(1.21) implies that electrons have physical existence only when the circumference of the permissible orbit is an integral multiple of the matter wave wavelength. This directly takes us to the Postulates of Neil Bohr.

Rewriting Eq.(1.15),

$I\omega = n\hbar = nh/(2\pi)$ 1.22

Where I = moment of inertia of the orbiting electron = mr^2 ;

ω = orbital angular velocity = v/r ;

Substituting the values of I and ω in Eq.(1.22) we obtain the following Equation:

$m.v.r = nh/(2\pi)$ 1.23

Reshuffling the terms we get:

$2.\pi.r = n.h/(mv_n) = n.h/(p_n) = n.\lambda_n$

Thus the first Postulate of Neil Bohr is a restatement of the fact that electron is a real material particle as well as it is a matter wave and hence electron's permitted orbits are radiationless and support the standing wave pattern only.

Neil Bohr inadvertently corroborated the wave nature of electrons..

When we are working at macroscopic level that is at cm and meter level, then Newtononian Mechanics and Maxwell's Electromagnetic Theory suffice for neutral and charged particles. But when we are working at atomic scale i.e. at Angstrom scale then we have to invoke Quantum Mechanics.

When we increase the scale from Angstrom to meter then Quantum Mechanics results correspond to those of Classical Mechanics. This known as Correspondence Principle.

SSPD_Chapter1_Part 7_Hydrogen Atom Analysis

SSPD_Chapter 1_ Part 7 gives the proof of Bohr's Quantum Theory of Atom postulates and lays the ground for giving a rigorous theoretical formulation of the line spectra of Hydrogen Atom and its corroboration with experimental values.

SSPD_CHAPTER 1_Part 7_ ANALYSIS OF HYDROGEN ATOM.

Consider a Hydrogen Atom where an electron is orbiting around the proton in circular path.

From Neil Bohr’s first postulate:

$m_e.v.r = n(h/(2\pi))$ 1.23

where n is the principal quantum number and an integral number.

Squaring both sides of Eq.(1.23) and reshuffling the terms,

$v^2 = n^2.\hbar^2/(m_e^2.r^2)$ 1.24

From rotational motion we know that:

Centripetal Force = $m_e.v^2/r$ = electrostatic force between electron and proton. = $q^2/(4\pi\varepsilon_0.r^2)$;

Therefore **$v^2 = q^2/(4\pi\varepsilon_0.r.m_e)$ 1.25**

Dividing Eq(1.25) by Eq.(1.24) we get:

$q^2/(4\pi\varepsilon_0 . r . m_e) = n^2 . \hbar^2/(m_e^2 . r^2)$

Simplifying the expression:

$r_n = n^2 . h^2 . \varepsilon_0 /(m_e . q^2 . \pi)$ = the Bohr Radius of nth Orbit; 1.26

$h = 6.62 \times 10^{-34}$ (J-sec) = Plank's Constant;

$m_e = 9.1 \times 10^{-31}$Kg;

$q = 1.6 \times 10^{-19}$ Coul;

ε_0 = absolute permittivity = $(1/(36\pi \times 10^9))$ F/m;

ε_r = relative permittivity of vacuum = 1;

$a_0 = \hbar^2/(m_e . q^2) = 0.4756$ = Bohr Radius;

By Dimensional Balance the correctness of the Equation(1.26) is established.

Eq.(1.26) tells us that only spheres of orbital radius $r_n = n^2 . h^2 . \varepsilon_0/(m_e . q^2 . \pi)$ are permitted.

The radius of the first and innermost spherical orbit = 0.529Å;

The radius of the second spherical orbit = 2.116 Å because n=2;

The radius of the third spherical orbit = 4.761 Å because $n=3$;

Thus we see that only a given number of spherical orbits having Bohr radii are permitted around the nucleus of an Atom. Also in these permissible orbits the standing wave boundary conditions are fulfilled as shown in Fig.(1.14)

i.e. $\mathbf{2\pi\ r\ n = n.\lambda\ n}$ **1.27**

From Eq.(1.26) $\mathbf{rn = n2.0.529}$**Å1.28**

Substituting Eq.(1.28) in Eq.(1.27)

$2\pi.n_2.\ 0.529$ Å $= n.\lambda_n$

Simplifying the expression:

$\mathbf{\lambda\ n = (n.\pi.1.058)}$ **Å 1.29**

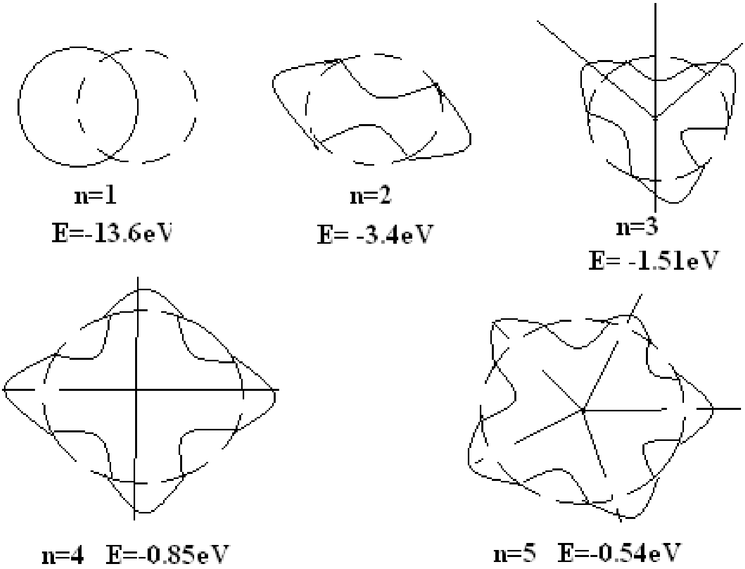


Figure 1.14 Standing Wave Pattern in permitted spherical orbits around the nucleus of Hydrogen Atom.

Figure 1.14. Standing Wave Pattern in permitted spherical orbits around the nucleus of an Atom.

From the two Equations (1.28) and (1.29), the radii of the permissible spherical orbits as well as the wavelength of the standing wave patterns are calculated as shown in Table(1.2).

Table 1. 2. The radii of the orbits and the wavelength of the standing wave pattern in first three permissible orbits.

n	r_n (Å)	λ (Å)
1	0.529	3.32
2	2.116	6.6476
3	4.76	9.97

The bounded electrons in an Atom are in stationary and non-radiating orbits *unlike* Betatron.

The bounded electrons orbital radii are quantized as shown by Eq.(1.28) as well as the energy of these bounded electrons in the spherical potential well are also quantized as we will show in the following derivation.

Total Energy = Kinetic Energy + Potential Well

$$E = (1/2)m_e v^2 - (q_e . q_p)/(4\pi\varepsilon_0 . r_n) \quad 1.30$$

In bringing a positive charge, work will be done against the repulsive field due the positive nucleus and hence potential energy will increase as we bring the positive charge nearer and nearer the nucleus. So a positive charge sees a potential hump at the positive nucleus. By similar logic , electron which is a negative charge sees a potential well at the positive nucleus hence potential energy is more and more negative as it falls deeper and deeper into the potential well.

$$E = (1/2)m_e v^2 - (q^2)/(4\pi\varepsilon_0 . r_n) \quad 1.30$$

From Eq.(1.25), $m_e.v^2 = q^2/(4\pi\varepsilon_0.r_n)$

Substituting the above expression in Eq.(1.30)

$E = -(1/2)m_ev^2$ 1.31

In Betatron, the Lorentz **[Appendix XXIX]** force provides the centripetal force to balance the centrifugal force created by the orbiting electron. Lorentz force is the motion created by left hand rule when a wire carrying current interacts with the transverse magnetic field. Lorentz force is given by Bqv where B is the transverse magnetic field and v is the linear velocity of the charge.

Therefore the equation of motion is:

Centripetal force = Bqv and centrifugal force = $(mv^2)/r$

Therefore **$(1/r)m_ev^2 = Bqv$ 1.32**

In Betatron, the orbiting electron is unbounded and is orbiting at macroscopic level hence it can be analyzed by the classical theory. Here the orbital radius can assume a continuum of values hence when electron emits synchrotron radiation it traverses a collapsing spiral path towards the center. When the electron’s energy is boosted by an oscillating electric field then electron traverses an outward expanding spiral path. Both these conditions once of emission and the other of absorption is shown in Fig(1.15).

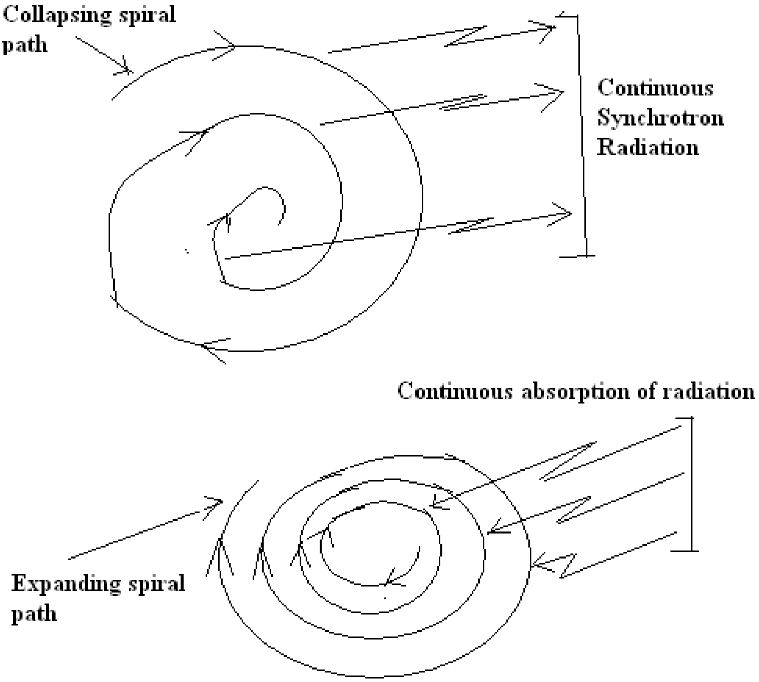

Collapsing spiral path
Continuous Synchrotron Radiation
Continuous absorption of radiation
Expanding spiral path

Figure 1.15 Collapsing and expanding spiral path of electron in Betatron

Figure 1.15 The inward collapsing spiral path of electron when emitting synchrotron radiation and the outward expanding spiral path of electron when boosted by alternating electric field.

This classical scenario is applicable to particles traversing circular paths at macroscopic scales. This is applicable to Betatron as well as to Synchrotron. But this picture completely fails at the Atomic scale.

For a Hydrogen Atom from Eq.(1.25) and (1.31) we obtain the following equation:

$E = -q^2/(8\pi\varepsilon_0 r_n)$

Multiplying numerator and denominator by r_n on RHS we get:

$E = -[q^2/(8\pi\varepsilon_0 r_n^2)] \times r_n$

Substituting the value of r_n from Eq.(1.26) we get:

$E = -[q^2/(8\pi\varepsilon_0 r_n^2)] \times (n^2.h^2.\varepsilon_0/(m_e.q^2.\pi))$

Simplifying the expression we get:

$$E = -(n^2/(2m_e.r_n^2))(\hbar^2) \quad 1.33$$

For the bounded electron , the orbital radii are quantized hence total energy in the permissible orbits are also quantized. Substituting Eq.(1.26) in Eq.(1.33) we obtain:

$$E = -(1/(32\pi^2)(m_e q^4/(\hbar^2 \varepsilon_0^2 n^2))\text{Joules} \quad 1.34$$

Putting Eq.(1.26) in Eq.(1.34) and rearranging the terms we get;

$E \times r_n = -(q^2/(8\pi\varepsilon_0))$

Substituting the value of

$r_n = n_2 \times (0.529\text{Å})$ in the above Equation we get the following Eq. for Hydrogen Atom:

$E = -\,(q_2/(8\pi\varepsilon_0 n_2 \times 0.529\text{Å}))$

Substituting the values:

E = -(13.6eV/n 2) 1.35

In Figure(1.16), the permissible energy states for an electron in a hydrogen potential well is illustrated.

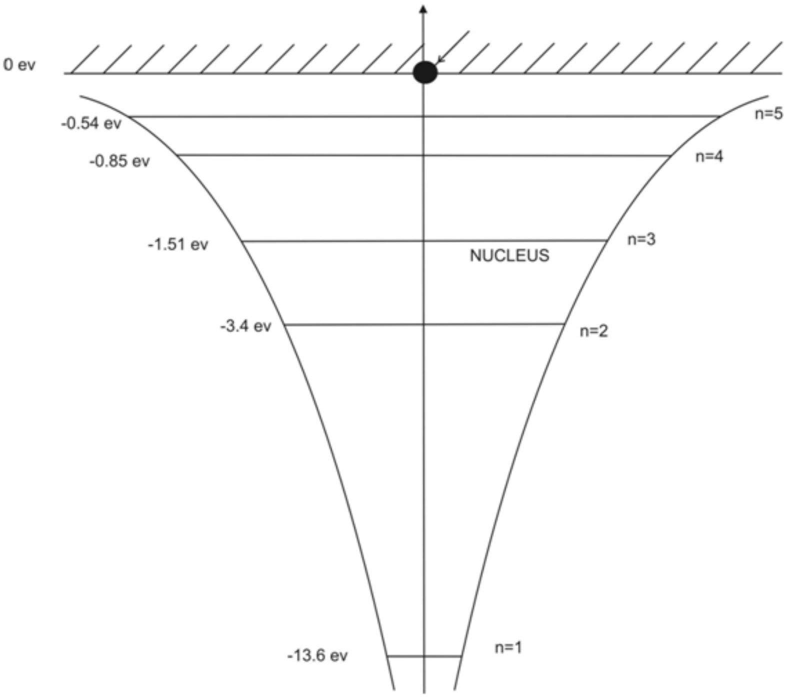


Figure 1.16. The quantized energy states of an electron in a hydrogen potential well.

Table(1.3) Discrete Energy States of Electron in Hydrogen Atom.

n	**E n**
1	13.6 eV
2	3.4eV
3	1.5eV
4	-0.85eV

In Hydrogen atom the electron is 13.6eV below zero hence we say an electron having principal quantum number n=1 is in a potential well -13.6eV deep. Therefore Ionization Energy of Hydrogen is 13.6eV.

Two Hydrogen atoms when brought together form a very strong covalent bond by pairing the two outer electrons and completing the octave condition of the orbit corresponding to n=1.

What this means that n=1 orbit or K Shell can accommodate only two electrons to completely fill up the available states in K Shell. Each Atom completes the octave by sharing each other’s electron for 50% of the time and thereby achieve the Gibb’s free Energy Minima. This Minima ensures a stable configuration to the Hydrogen molecule.

This energy minima is - 4.5eV. In other words the binding energy of the Hydrogen molecule is 4.5eV.

If Hydrogen gas is heated to $T_1 = 15.6 \times 10^4$ Kelvin , the thermal energy imparted is $kT_1 = 13.4$eV. Since this exceeds 4.5eV hence Hydrogen Molecule dissociates into two independent Hydrogen Atoms. And there is sufficient energy for the first ionization of Hydrogen Atom. When electron from K Shell is removed from the atom then we have First Ionization. If electron is removed from L Shell then we have Second Ionization and so on. We will see later that in the beginning when Big Bang took place the matter was in gluonic state and then in plasma state. Radiation was coupled with Matter. After 300,000 years when temperature fell to 4000K then plasma neutralized and radiation decoupled from matter and as Universe expanded this radiation went on cooling and today it appears as 3Kelvin relic radiation or as Cosmic Microwave Back Ground (CMB)Radiation and it carries the footprint of matter distribution as it was at the time of decoupling i.e. 300,000 years after the Big Bang. This CMB Radiation has hot spots meaning by there are clumping of matter in certain parts of space and these matter clumps acted as the nucleus of Galaxy Growth. CMB Radiation was isotropic then matter density would be homogeneous through out the Universe and there would have been no evolution of Galaxies, stars and planets.

At the Angstrom scale, only Quantum Mechanics correctly explains the behaviour of a bound electron in an Atomic potential well. This is confirmed by experimental observations. Experimentally it has been found that every element has its characteristic line spectrum. The line spectra could be of emissive type or absorption type.

When an excited element comes down to the ground state, it gives off certain characteristic wavelengths . This is called the emissive line spectra.

When a ground element intercepts an incident light, it absorbs certain characteristic wavelengths creating dark fringes in the spectrum of the resulting light. This is described as the absorption spectra of the given element.

During the period 1885-1908, the characteristic spectra was described by the following equation:

$$N = \nu/c = 1/\lambda = R(1/n_2^2 - 1/n_1^2) \quad 1.36$$

Where ν frequency of the transverse electro-magnetic wave;

Wavelength = λ;

Velocity of light = c;

Wave number = N;

Rydberg constant = 109677.6 cm^{-1};

n_1 = 1,2,3…..;

n_2 = 2,3,4,……;

The prominent Scientists associated with spectrum research at the time were Ritz, Rydberg and Barmer **[Appendix XXX]** .

Neil Bohr was able to give a theoretical explanation for Eq.(1.36) and theoretically derive Rydberg Constant.

SSPD_Chapter1_Part 7continued_Theoretical Formulation of Rydberg Constant

This is the continuation of Chapter 1_Part 7. This gives the theoretical explanation of the five distinct series of LINE SPECTRA experimentally observed in Hydrogen Gas.

1.6.3.2. THEORETICAL DERIVATION OF RYDBERG CONSTANT.

When a photon of energy packet hν is incident on an atom of a given element , it excites an electron from ground state n_1 to the excited state n_2 in a way so as to satisfy the following condition:

$\Delta E = E_{n2} - E_{n1} = h\nu$;

From Eq.(1.35) the condition to be satisfied is:

$h\nu = (-13.6/n_2^2 + 13.6/n_1^2)$ eV;

or $N = 1/\lambda = \nu/c = (13.6\text{eV}/hc)(-1/n_2^2 + 1/n_1^2)$ 1.37

By comparing Eq(1.36) and Eq(1.37) we arrive at the numerical value of Rydberg Constant R to be:

$R = (13.6\text{eV}/hc)$ 1.38

Substituting the numerical values of h and c in Eq.(1.38) we obtain :

$$\mathbf{R = 1.096 \times 10\ 5\ cm\ {-1}\ \ 1.39}$$

The theoretical value of Rydberg Constant is close to the experimental value as given by Eq.(1.36).

Theoretical derivation of Rydberg Constant is considered as a definite vindication of the Quantum Theory of Atoms as proposed by Neil Bohr.

Based on this interpretation we get the underlying explanation for the various line spectra exhibited by Hydrogen Atom namely Lyman Series or K series, Balmar Series or L series, Paschen series or M series, Brackett series or N series and Pfund series or O series. The excitation states which create the five series of line spectra are shown in Fig.(1.17).

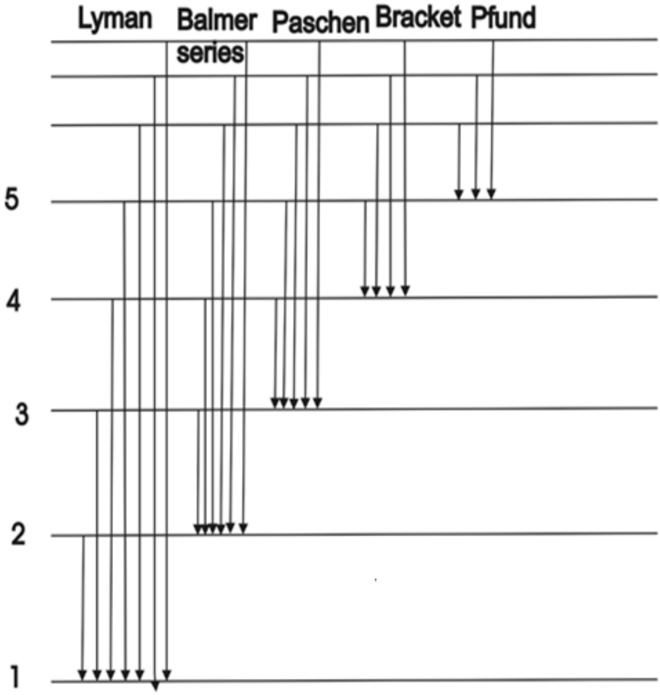


Fig.(1.17) Basic underlying cause for the generation of the five line spectra of the Hydrogen Atom

Fig.(1.17) Basic underlying cause for the generation of the five line spectra of Hydrogen Atom.

Based on these line spectra the shell corresponding to principal quantum number n = 1 that is the innermost orbit is referred to as K Shell. Similarly the subsequent orbits corresponding to n = 2,3,4,5 are called L, M, N, O Shells respectively.

Atomic Energy Exchanges and Physics of Gas Discharge Tubes.

This is the continuation of Chapter 1_Part 7 where we describe the resonant transition , fluorescence, phosphorescence and we bring out the distinction between spontaneous and stimulated emission.

SSPD_Chapter1_Part7continued_ ATOMIC ENERGY EXCHANGES AND PHYSICS OF GAS DISCHARGE TUBES.

Atomic Energy Exchanges are governed by resonance phenomena. If the incident photon energy is approximately equal to the transition energy then photon is said to resonate with transition and the excitation of the electron at the given transition is favored.

For 1 electron atom:

$$\mathbf{P(transition) = (Sin\Theta/\Theta)^{2} = Sinc^{2}(\Theta) = Sin^{2}\,[\pi(\nu\ ko - \nu)t]/[\pi(\nu\ ko - \nu)t]^{2}} \quad \mathbf{1.40}$$

Where P(transition) is the probability of transition from ground state to the excited state k.

(*Zambuto, Mauro: Handbook of Electrical and Computer Engineering (ed. Sheldon S.L. Chang) , Vol.1, pp. 597- 623, John Wiley & Sons, Inc, New York, 1994;*

Leighton, Robert A. Principles of Modern Physics, McGraw Hill, New York, 1954)

Transition Energy from ground state to excited state $= h\, \nu_{ko} = (E_n - E_0)$;

Incident Photon Energy $= h\, \nu$;

$t =$ duration of interaction;

For reasonable interaction duration, only incident photon with frequency $\nu = \nu_{ko}$ will induce excitation. Therefore only resonant photons are absorbed. To the rest of the photons, the atom appears to be transparent.

Probability of Absorption is sharply peaked at resonance as shown in Fig.(1.18)

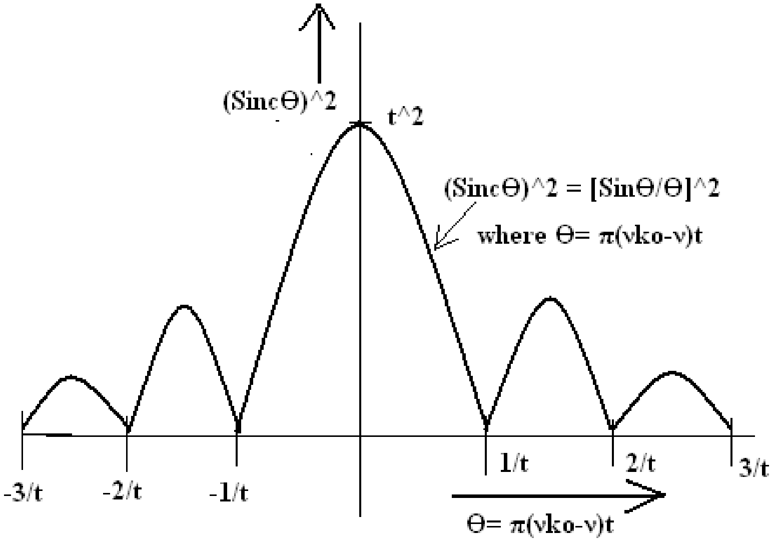


Figure 1.18. Probability of absorption is given by (SincΘ)^2 where Θ= π(vko-v)t.

Fig.(1.18) Probability of Absorption is given by (SincΘ) 2 where Θ = π(ν ko -ν)t

Spontaneous Emission Phenomena.:

When an excited atom returns to ground state in the natural course of events and in the process emits a photon equal to the transition energy , we say spontaneous emission has taken place. This process of spontaneous emission is also known as relaxation.

When relaxation is in nanosecond, we have fluorescence effect. In fluorescence the excited atom relaxes by emitting a lower frequency photon as compared to the higher frequency absorbed. The difference between absorbed and emitted is lost due to collisions with other molecules. The energy absorbed resides in the molecule as rotational energy, vibrational energy and as excited electron energy. The lowering of emitted photon takes place due to the loss of a part of the vibrational energy. We will have a detailed discussion on this topic in the last section, specific heat section, of this Chapter.

When relaxation is in micro and millisecond, we have phosphorescence effect. The photo emission is delayed and prolonged. Here the absorption of photon causes the transition from ground singlet state to excited singlet state. In singlet state total spin quantum number is $S = 0$. Due to collisions with other molecules, the excited singlet state loses a part of vibrational energy and suffers a radiationless transition to excited triplet state $(S = 1)$. The energy difference between the vibrational energy states is lost as the thermal energy of the gas molecules. Now a radiative transition from excited triplet state to ground singlet state is forbidden by SELECTION RULE [Appendix XXXI A]. Though forbidden nevertheless after some time the excited triplet molecule does relax to the ground singlet state thus giving rise to a delayed and prolonged emission. This is termed as Phosphorescence. The delayed emission may occur after minutes or even after hours depending upon the composition of phosphorescence screen.

This phosphorescence effect is used in Oscilloscope screens where we want persistence of the image.

In mercury lamps the fluorescence effect is used. In a mercury tube, a vacuum tube contains a drop or two of molten mercury along with inert Argon gas. This is vaporized and a large voltage is applied to cause an electric discharge in mercury vapour. In the process the mercury atoms are ionized. Almost instantaneously ionized atoms neutralize by the process of spontaneous emission and emit Ultra Violet light which is invisible. The UV light strikes the walls of the tube which is coated with fluorescence material called phosphor. This coating gets excited by the absorption of UV light and immediately relaxes to the ground state by emitting soft white light. Here relaxation time is of the order of nanosecond. Hence Mercury Tube is also called Fluorescence Tube. Flourescence is much more efficient than Incandescence used in Incandescent Lamps invented by Thomas Alva Edison **[Appendix XXXII]**.

In a similar fashion the discharge in various gases cause spontaneous emission of various colored light.

Mercury Discharge Physics:

Hg(vapor) + collision energy(4.9eV) = Hg* (ionized mercury atoms);

(Collision Energy is caused by highly accelerated ionized particles in presence of critical electric field. The ionized particles are caused due to cosmic particles striking the mercury vapor atoms).

Hg* = Hg + 4.91eV

And λ = 1.24/(4.91eV) = 0.252μm (*Ultra-violet light*);

1st ionization energy (energy required for stripping the outermost electrons) = 2.00eV.

Sodium Vapor Discharge Lamp:

Na + collision energy (2.85eV) = Na*

Na* = Na + 2.85eV photon;

And λ = 1.24/(2.85eV) = 0.435μm (*Yellow light*);

First Ionization Energy : E_1 = 495 KJ/MOLE; (0.93eV)

Second Ionization Energy : E_2 = 4560 KJ/MOLE;

Third Ionization Energy : E_3 = 6900 KJ/MOLE;

Fourth Ionization Energy : E_4 = 9540 KJ/MOLE;

Fifth Ionization Energy : E_5 = 13400 KJ/MOLE;

Sixth Ionization Energy : E_6 = 16600 KJ/MOLE;

In first ionization the outermost electrons are being ripped apart from the atom. In second ionization, the second outermost electrons are being ripped apart from the atom. And so on.

Neon Discharge Lamp:

Ne + collision energy(10.6eV) = Ne* Ne* = Ne + *Orange-red light*;

$E_1 = 2080$, $E_2 = 3950$, $E_3 = 6120$, $E_4 = 9370$, $E_5 = 12200$, $E_6 = 15000$ (KJ/MOLE);

First ionization energy = 4.84eV

Argon Discharge Lamp:

Ar + collision energy(7.7eV) = Ar*

Ar* = Ar + *Bluish Purple*;

$E_1 = 1521$, $E_2 = 2666$, $E_3 = 3900$, $E_4 = 5770$, $E_5 = 7240$, $E_6 = 8800$ (KJ/MOLE);

First ionization energy = 3.2eV

Krypton Discharge Lamp:

Kr + 6.8eV(1400 KJ/Mole) = Kr*

Kr* = Kr + *White light*

First ionization energy $=$ 2.94eV

Helium Discharge Lamp :

He $+$12.3eV(2373 KJ/MOLE) $=$He *

He* $=$He $+$?

E_2 $=$5248 KJ/MOLE

Hydrogen Discharge Lamp:

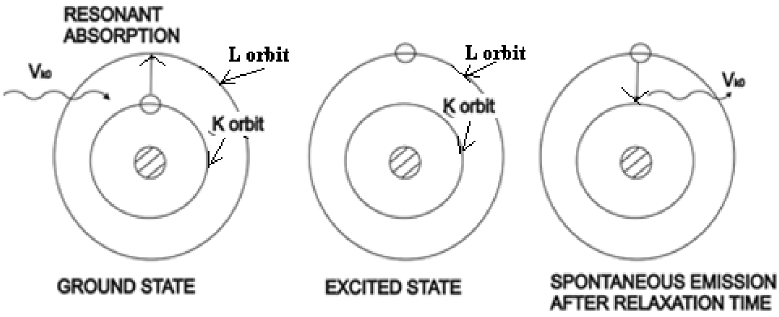
H $+$ 7.18eV $=$ H*

Stimulated Emission Phenomena:

Stimulated Emission is induced emission when incident photon matches the transition energy . This effect is used in LASER(Light Amplification by Stimulated Emission Radiation).

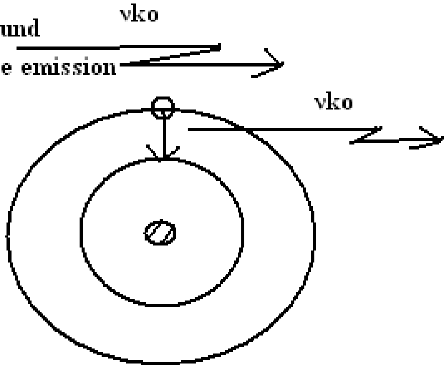
In Fig.(1.19) the spontaneous emission is shown.

In Fig.(1.20) the stimulated mission is shown.



Fig(1.19) The mechanism of Spontaneous Emission

Fig.(1.19). The mechanism of Spontaneous Emission.



The passage of a Photon vko causes the transition to Ground State thereby stimulating the emission of a photon of identical frequency and phase.

This is called STIMULATED RADIATIVE TRANSITION.

Figure 1.20. The mechanism of Stimulated Emission.

Fig(1.20). The mechanism of Stimulated Emission.

SSPD_Chapter1_Part 8_Electron in a Potential well_Formulation of Schrodinger Equation. SSPD_Chapter 1_ Part 8 deals with Electron in a Potential Well. To solve this problem we first write the Hamiltonian and formulate the Schrodinger Equation.

SSPD_Chapter 1_Part 8_ ELECTRON IN 1-D POTENTIAL WELL_Formulation of Schrodinger Equation.

One Dimensional Potential Well can be of three types:

1. 1-D Potential Well of infinite height;
2. 1-D Potential Well of finite height;
3. 1-D Potential Well of finite height and finite thickness;

Before analyzing electron in a potential well we will give the physical significance of matter wave.

Like electromagnetic waves, matter wave is a progressive or traveling harmonic wave.

For mathematical simplicity we will deal with waves having plane wavefront. These are called Plane Waves.

There can be traveling waves of spherical wavefront or cylindrical wavefront and these are known as spherical waves or cylindrical waves respectively.

In fact a spherical wave or cylindrical wave becomes a plane wave after they have traveled for a very long distance. As long as radius of curvature is finite we have cylindrical or spherical curvature and curvature is defined as 1/r where r is the radius of curvature. As soon as radius of curvature is infinite as it would be if the wave has traveled for a very long distance , the curvature becomes zero and the wave assumes a plane wavefront and it is called Plane Wave.

In Fig.(1.21) a Plane Transverse Electro-Magnetic Wave is shown.

$$\mathbf{E(z,t) = E\ x0\ Exp\,[j(k\ z\ .z - \omega.t)]}$$

$$\mathbf{H(z,t) = H\ y0\ Exp\ \ [j(k\ z\ .z - \omega.t)]\ \ 1.41}$$

Where $k_z = 2\pi/\lambda_z$ = wave propogation vector or wave number;

And $\omega = 2\,\pi\,\nu$ = circular frequency;

A wave front is defined as front on which all points have same phase. If this front is plane we say we have a plane wave. If the front is cylindrical, we have cylindrical waves and if the wave front is spherical we have spherical waves.

The wave front moves with a velocity . This wave front velocity is the velocity of wave propogation.

i.e. $k_z.z - \omega.t = \text{constant} = K$

Taking the time derivative we obtain:

$k_z.\partial z/\partial t - \omega. = 0$

rearranging the terms we get: .

$\partial z/\partial t = \omega/ k_z = \lambda_z . \nu$ = velocity of the progressive wave front.

A phase term of the form $(k_z.z - \omega.t)$ gives a forward traveling wave

whereas a phase term of the form$(k_z.z + \omega.t)$ gives a backward traveling wave.

In absolute vacuum Electromagnetic Wave travels with velocity of light $c=3\times10^8$m/sec.

In absolute vacuum, $\omega/k_z = \lambda_z.\nu = c = 3\times10^8$m/sec (velocity of light, in absolute vacuum,is suppose to be an invariant quantity i.e. invariant with respect to the frame of reference **[Lorentz Invariance or Gauge Invariance – Appendix XXIX]** . This fact was established by Albert Abraham Michelson(1852-1931) and Edward Williams Morley(1838-1923) in their celebrated experiment known as Michelson-Morley experiment **[Appendix XXXIII]** . This is called Special Theory of Relativity which was propounded by Einstein in 1905).The Eq. (1.41) is the solution of the wave equation given in **Appendix(XXXIV)**.

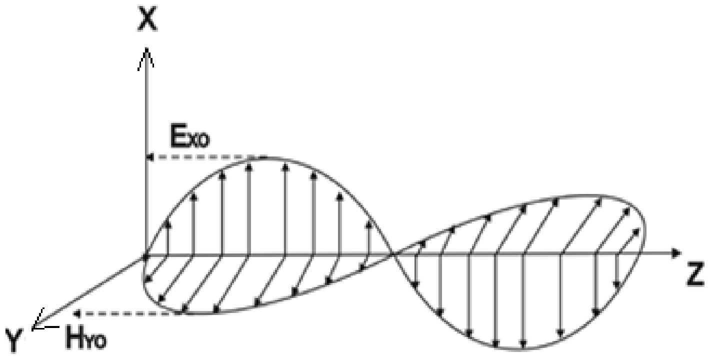


Figure 1.21. A progressive Electro-Magnetic Wave.

Fig(1.21) A Progressive Plane Electro-Magnetic Wave.

An electromagnetic wave is a transverse wave because the oscillation is transverse to the direction of propagation. Electric field is oscillating in X direction and Magnetic Field is oscillating in Y direction whereas Z axis is the direction of propagation. Here X and Y axes are transverse to Z-axis. Maxwell equations constrain Electric field and Magnetic Field to be perpendicular to each other and also to the direction of propagation.

In contrast sound wave is a longitudinal wave because the medium particles oscillate in the same direction as the direction of propagation.

In an identical manner Schrodinger **[Appendix XXXV]** proposed the progressive matter wave:

In 1926 Schrodinger stated the Hamiltonian Equation:

$H. \psi = E. \psi$

where H is the Hamiltonian **[Appendix XXXVI]** Energy Operator $= p_2/(2m) + V(r)$;

and E is the total energy;

Using this Schrodinger Equation, solutions are obtained for an electron in a Hydrogen atom and the four quantum numbers are predicted namely Principal Quantum Number(n), Azimuthial Quantum Number (l), Magnetic Quantum Number (m) and Spin Quantum Number (s). These quantum numbers correctly explain the experimental results.

Still there was no direct proof for the existence of matter wave.

The mathematical form of the solution of Schrodinger Equation is :

$\psi(z,t) = \psi 0 .Exp[j(k z .z - \omega.t)]$ 1.42

Where $k_z = 2\pi/\lambda_z$ = wave propogation vector or wave number;

And $\omega = 2\pi\nu$ = circular frequency;

And ψ = probability amplitude;

And $\psi . \psi^* = |\psi|^2$ = probability density = probability of finding the particle per unit volume;

$|\psi|^2 dV$ = probability of finding the particle in an elemental volume.

If $|\psi|^2 dV$ is integrated over the total volume where the particle is likely to be found then the integral will come out to be unity. This means the particle is certainly present in the given volume over which the integration has been carried out.

If the integral is zero then it means it is impossible to find the particle in the given volume.

If integral is between 0 and 1 say 0.5 that means if hundred observations are made within the given volume then in fifty observations the particle is likely to be detected.

If integral is 0.25 then it implies that out of 100 observations 25 times the particle will be observed.

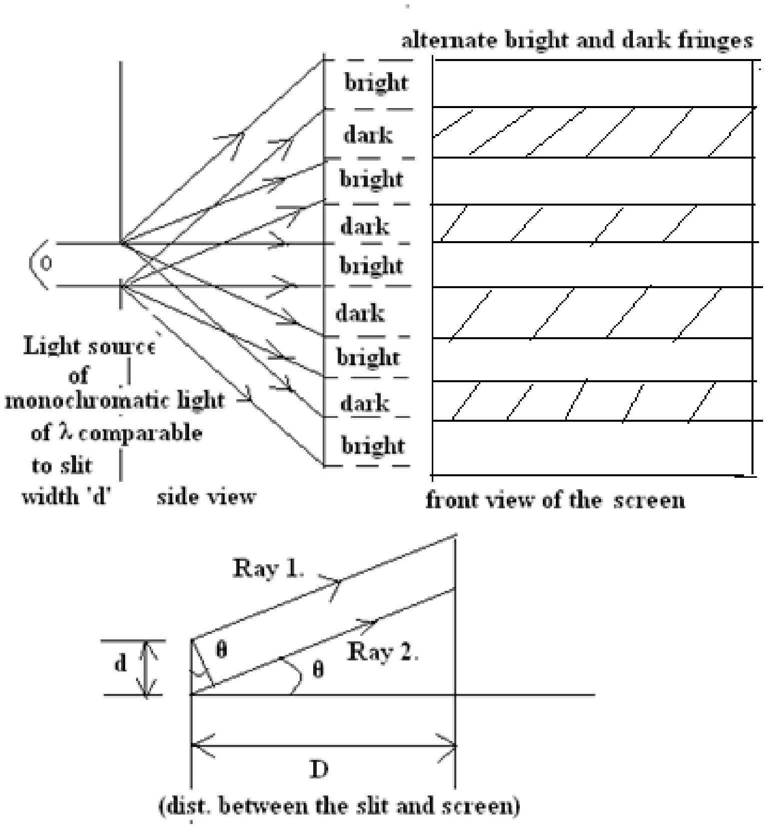
Therefore closed $\int_V [|\psi|^2 dV]$ = probability of finding the particle in volume V;

Therefore $\int \mathbf{V}\ [|\ \psi\ |\mathbf{2dV}] = \mathbf{1}$ implies certainty of finding the particle in V **1.43**

And $\int \mathbf{V}\ [|\ \psi\ |\mathbf{2dV}] < \mathbf{1}$ implies less than 100% probability of finding the particle in volume V **1.44**

In spite of this correctness there were no direct experimental proofs for the matter wave though a THOUGHT EXPERIMENT had been proposed in early 70’s.

In 1985 this thought experiment was practically performed as shown in Figure(1.22).


alternate bright and dark fringes
bright
dark
bright
dark
bright
dark
bright
dark
bright
Light source of monochromatic light of λ comparable to slit width 'd'
side view
front view of the screen
Ray 1.
Ray 2.
θ
θ
d
D
(dist. between the slit and screen)

Bright Fringe is obtained when there is constructive interfernce between Ray 1 and Ray 2. Condition for constructive interference is that path difference between Ray 1 and Ray 2 should be n λ.

Figure 1.22. Single Slit Diffraction Experiment.

Fig(1.22) Single slit light diffraction experiment.

Path length of ray 1 is = D/CosΘ

Path length of ray 2 is = D/CosΘ + d.SinΘ

Therefore the path difference = d.SinΘ

When d.SinΘ = nλ then we have a bright fringe and

When d.SinΘ = (n + ½)λ then we obtain a dark fringe. Thus Single Slit Diffractin pattern of light consists of alternate bright and dark fringes.

If instead of light we use a very weak source of electrons then instead of bright and dark fringes we obtain the high and low probability of finding the electrons on different parts of the screen.

For this experiment the electron source should be weak enough to pass singally or one by one through the single slit. As a detector the positively charged anode plate should be used and the whole diffraction set up should be installed in an evacuated chamber as shown in Fig(1.23). The strength of current picked up from different parts of the screen is directly proportional to the probability of incidence of the incident electrons. In place of bright fringe we will obtain a large current and in place of dark fringe we will obtain a small current.

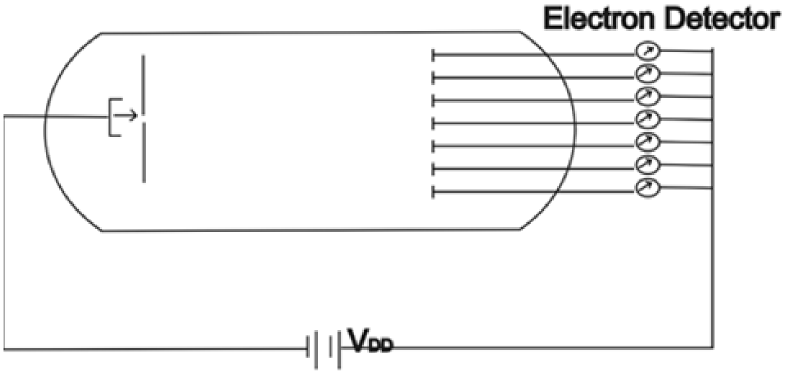


Fig1.23 Electron Diffraction Set Up

Fig(1.23) Electron Diffraction Set Up.

After giving a physical interpretation of the matter wave we will come to the statement of Schrodinger Equation.

Position (z) and momentum in z direction (p_z) are defined as canonical conjugate variables.

Similarly time (t) and energy (E) are also defined as canonical conjugate variables.

Momentum p_z is defined as operator $i\hbar\partial/\partial z$ and

Energy E is defined as $-i\hbar\partial/\partial t$.

Schrodinger used

Hamiltonian Operaror H = Kinetic Energy Operator + Potential Energy Operator.

Therefore **$H = p2/(2m) + V(z)$ 1.45**

When Operator H operates upon a matter wave then it yields the Eigen value of the energy of the matter wave.

Therefore **$H\psi = E\psi$ 1.46**

Therefore $[p^2/(2m) + V(z)]\ \psi = E\psi$

Substituting p operator we get:

$[-(\hbar^2/2m)\ \partial^2/\partial x^2 + V(z)]\psi = E\psi$

Therefore $-(\hbar^2/2m)\ \partial^2 \psi /\partial z^2 = [E-V(x)]\ \psi$

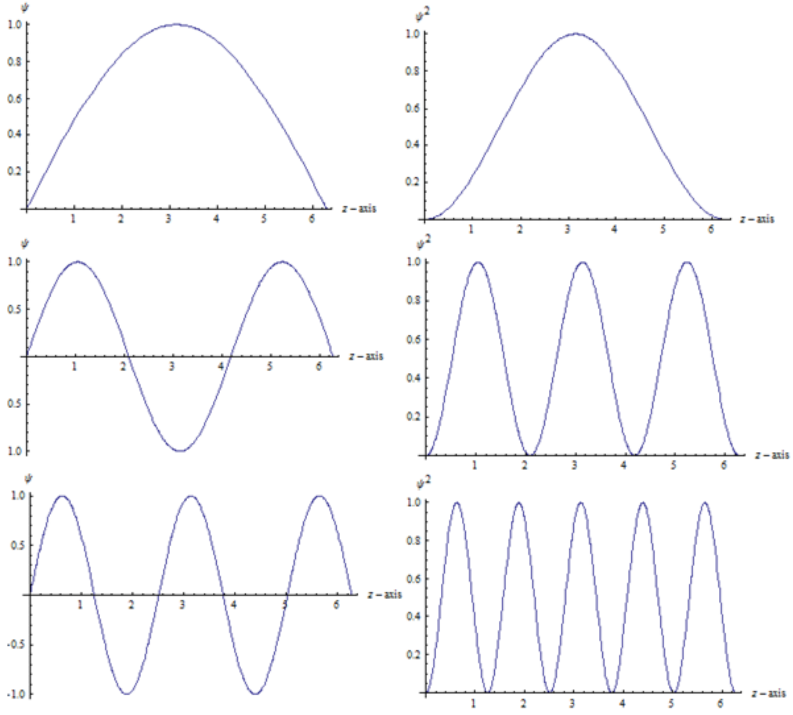
Or $(\hbar\ 2\ /2m)\ \partial\ 2\ \psi\ /\partial z\ 2 + [E-V(z)]\ \psi = 0$ 1.47

This is a second order linear differential equation known as the Schrodinger Equation and this has a standard procedure for its solution.

SSPD_Chapter1_Part 8continued_Electron in an Infinite Potential Well_Solution of Schrodinger Equation.

This is the continuation of Chapter 1_Part 8. This gives the solution of Schrodinger Equation for an electron in an 1-D Infinite Potential Well.This is analogous to the analysis of an electron in the ground state of hydrogen atom except that for a hydrogen atom the Schrodinger Equation will have to be solved for spherical coordinates. The problem of Hydrogen Atom will be taken up in a latter chapter.

SSPD_Chapter 1_Part8 continued_ SOLUTION OF SCHRODINGER EQUATION FOR AN ELECTRON IN AN INFINITE POTENTIAL WELL.



Fig(1.24) Electron in 1-D infinite potential well. Probability Amplitude ψ(z,t) and probability density |ψ(z,t)|ˆ 2 are respectively plotted along the spatial axis-z axis.

Writing the Schrodinger Equation **[Appendix XXIX]** for an electron in an infinite potential well where V(z) = 0 we get:

$$(\hbar^2/2m)\,\partial^2\psi/\partial z^2 + E\psi = 0 \quad 1.48$$

$$\partial^2\psi/\partial z^2 + (2mE/\hbar^2)\psi = 0$$

According to Linear Operator Theory, second order linear differential equation has two solutions: the complementary solution also known as transient solution or the natural solution and the second solution is the particular solution also known as the steady state solution or the forced solution.

The total solution = transient solution + steady state solution.

The transient solution depends on the characteristic equation of the system. A second order linear system has second degree characteristic equation which has two roots and two arbitrary constants which need two boundary conditions.

The steady state solution depends on the forcing function or the driving function. If the forcing function is a harmonic function then the steady state solution is a sinusoidal solution. If the forcing function is a constant the steady state output is also a constant.

In Eq.(1.48) Right Hand Side is zero hence steady state solution is zero.

The characteristic equation is:

$D_2 + 2mE/\hbar_2 = 0$

Where D is the differential operator.

The roots are imaginary:

$D_1 = -i[\sqrt{(2mE)}]/\hbar$ and $D_2 = +i[\sqrt{(2mE)}]/\hbar$

Therefore the solution is harmonic. If the roots were real then the solution would be hyperbolic or exponential.

The solution of Eq.(1.48) is:

$$\boldsymbol{\psi = A.\mathrm{Exp}[+i\{(\sqrt{(2mE)})/\hbar\}x] + B.\mathrm{Exp}[-i\{(\sqrt{(2mE)})/\hbar\}x]};$$

From two boundary conditions the two arbitrary constants are determined.

At $x = 0$, $\psi(0) = 0$ and at $x = W$ (the width of the potential well), $\psi(W) = 0$.

From these two boundary conditions we obtain two simultaneous equations.

$$A + B = 0;$$

$$A.\mathrm{Exp}[+i\{(\sqrt{(2mE)})/\hbar\}W] + B.\mathrm{Exp}[-i\{(\sqrt{(2mE)})/\hbar\}W] = 0;$$

The determinant is:

$$\Delta = 1 \quad 1$$

$$\mathrm{Exp}[+i\{(\sqrt{(2mE)})/\hbar\}W] \quad \mathrm{Exp}[-i\{(\sqrt{(2mE)})/\hbar\}W]$$

Therefore $\Delta = \mathrm{Exp}[-i\{(\sqrt{(2mE)})/\hbar\}W] -$

Exp[+i{(√(2mE))/ ħ}W]

If these two simultaneous equations are to have non-trivial solutions then the determinant of the set of simultaneous equations should be zero.

Therefore Δ = Exp[-i{(√(2mE))/ ħ}W] - Exp[+i{(√(2mE))/ ħ}W] =0

Therefore Exp[-i{(√(2mE))/ ħ}W] = Exp[+i{(√(2mE))/ ħ}W]

Therefore Cos[{(√(2mE))/ ħ}W] –i Sin[{(√(2mE))/ ħ}W]

= Cos[{(√(2mE))/ ħ}W] +i Sin[{(√(2mE))/ ħ}W]

Or –i Sin[{(√(2mE))/ ħ}W] = i Sin[{(√(2mE))/ ħ}W]

This relationship is satisfied only when Sin[{(√(2mE))/ ħ}W]=0;

This requires {(√(2mE))/ ħ}W = 0 radians or

{(√(2mE))/ ħ}W = π radians or

{(√(2mE))/ ħ}W = n π radians where n= 0,1,2,3,4…….

Squaring both sides we get:

$\{(2mE)/\hbar^2\}W^2 = (n\pi)^2$

or $E = n^2 h^2/(8mW^2)$ **1.49**

Eq.(1.49) implies that the permissible energy states of an electron in an infinite potential well are quantized.

Why electrons cannot occupy a continuum energy states as they do in free space? The answer is the following:

We had assumed at the beginning of the analysis that $V(x) = 0$. This implies that potential energy is zero and electron possesses only Kinetic Energy.

Therefore total energy $E = p^2/(2m) = n^2h^2/(8mW^2)$

Therefore $p = (nh)/(2W)$ **1.50**

From de Broglie postulate: $\lambda = h/p = h/[(nh)/(2W)]$

To satisfy the standing wave condition in bounded space which an infinite 1-D potential well is, following boundary condition must be satisfied

$W = n\lambda/2$ **1.51**

Eq.(1.51) is the necessary condition for Standing Wave pattern. This standing wave pattern requirement causes the quantization of energy states.

Here we digress briefly to the chapter of light to fully understand the behavior of matter wave.

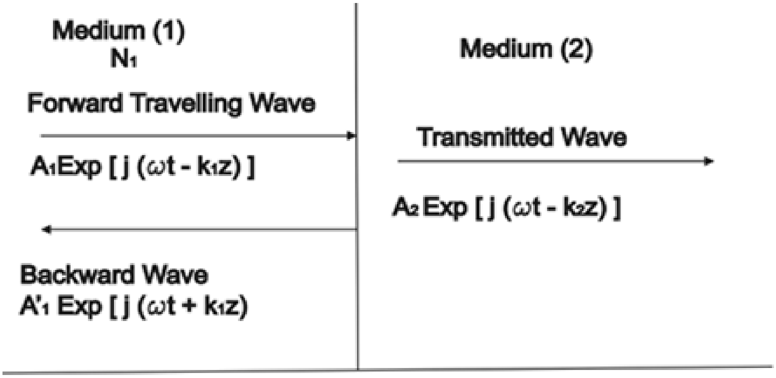


Figure 1.25. A plane optical wave incident normally on the interface of two different optical media.

Fig.(1.25) A plane wavefront light ray perpendicularly incident on an interface of two optical mediums.

Whenever light travels from one optical medium of refractive index n1 to the other optical medium of refractive index n2 , the incident wave Transverse Electromagnetic Wave (TEM) experiences partial reflection at the interface of the two media and partial transmission into the second medium.

Let us assume that medium 1 is absolute vacuum hence its refractive index = n1 = 1 and medium 2 is a solid medium of refractive index $n_2 = n$. The mathematical form of the incident wave, reflected wave and the transmitted wave is given in Fig(1.25).

Wave vector in medium 1 is $k_1 = 2\pi/\lambda_1$ and wave vector in medium 2 is $k_2 = 2\pi/\lambda_2$;

And ν **λ1 = c,** ν **λ2 = v; 1.52**

Therefore $c/v = \lambda_1 / \lambda_2 = n/1$;

Therefore **λ1 = n. λ2 1.53**

We know that if the second medium is metal, the incident light is totally reflected and the reflected light experiences a phase change of 180°.

Also at the interface $E_{incident} + E_{reflected} = 0$;

But if the second medium is dielectric then we have partial reflection and partial transmission and at the interface we have: $E_{incident} + E_{reflected} = E_{transmitted}$;

The incident wave is the forward wave:

E(z,t) = E xoincident Exp [j(k z1 z – ω.t)] ;

The reflected wave is the backward wave:

E(z,t) = E xoreflected Exp [j(k z1 z + ω.t)] ;

The transmitted wave is also moving in forward direction therefore it is:

$E(z,t) = E\ \mathrm{xotransmitted}\ \mathrm{Exp}[j(k\ z2\ z - \omega.t)]$; 1.54

Wave vectors have been defined in Eq.(1.52).and Eq.(1.53).

The incident forward and reflected backward wave interfere to form Standing Wave as they do on a mismatched transmission line. If a transmission line is not terminated in Characteristic Impedance then partial reflection takes place at the load and a partial standing wave pattern is formed on the transmission line. Standing Wave implies there is no transmission of energy. In case of metal there is total reflection. Hence we have 100% standing wave in medium 1 and there is no penetration of light into the metallic medium 2. Hence no transmission of light energy. For dielectric medium 2 , we have partial reflection hence only partial standing wave.

The abrupt change in the refractive index or the dielectric constant is referred to as step index change. In the same way abrupt voltage change is referred to as step voltage change.

Matter wave behaves in an analogous fashion at a step voltage change as a light wave behaves at step index change.

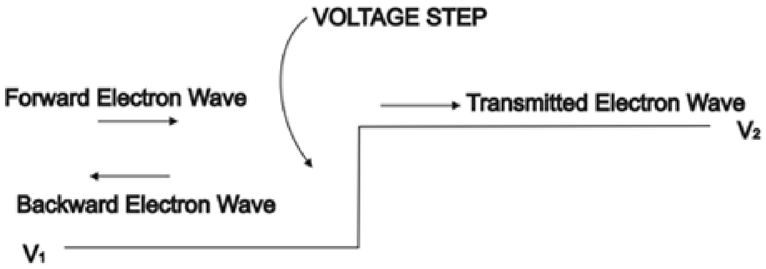


Figure 1.26. Reflection and Transmission of electron matter wave across step voltage change from V1 to V2.

Fig(1.26) Reflection and Transmission of electron matter wave across step voltage change from V 1 to V 2 .

For matter wave,

wave vector $k = 2\pi/\lambda = 2\pi p/h = [\sqrt{2m(E-V)}]/\hbar$;

Therefore the two wave vectors are:

$k_1 = [\sqrt{2m(E-V_1)}]/\hbar$ and $k_2 = [\sqrt{2m(E-V_2)}]/\hbar$;

The incident forward electron matter wave:

$\boldsymbol{\psi(z,t)_{\text{incident}} = \psi_{0\text{incident}} \operatorname{Exp}[j(\omega t - k_{z1} . z)]}$

The reflected backward electron wave:

$\boldsymbol{\psi(z,t)_{\text{reflected}} = \psi_{0\text{reflected}} \operatorname{Exp}[j(\omega t + k_{z1} . z)]}$ **1.55**

The transmitted forward electron wave:

$\psi(z,t)_{\text{transmitted}} = \psi_{0\text{transmitted}}\,\text{Exp}\,[j(\omega t- k_{z2} .z)]$ 1.56

At the interface or at the step,

$\psi_{0\text{incident}} - \psi_{0\text{reflected}} = \psi_{0\text{transmitted}}$ 1.57

Eq.(1.57) is the interface boundary condition.

Now we return to the original infinite potential well case.

Inside the potential well we have both the wave vectors $\pm i[\sqrt{2mE}/\hbar]$ hence we have both forward and backward traveling wave components.

Outside the well Schrodinger Equation is of the form:

$$\partial^2 \psi /\partial x^2 + (2m(E-V)/\hbar^2)\psi = 0$$

But outside the well, $V = \infty$ therefore the wave equation reduces to:

$$\partial^2 \psi /\partial x^2 - (2m\infty/\hbar^2)\psi = 0$$

The characteristic equation has real roots hence the solution is hyperbolic.

Therefore $\psi(z,t) = [A\text{Exp}(-k_2 .z) + B\text{Exp}(+k_2 .z)]\,\text{Exp}(i\omega t)$ 1.58

Where $k_2 = [\sqrt{2m\infty}]/\hbar]$

In Eq(1.58), the term $\text{Exp}(+k_2.z)$ is not admissible as it is exponentially growing function. Hence $B=0$ and the solution outside the well is:

$\psi(z,t) = [A\text{Exp}(-k_2 .z)]\,\text{Exp}(i\omega t)$ 1.59

But since k_2 = infinity implies the wave function does not exist outside the well. The transmitted wave outside the well almost instantaneously attenuates to zero since k_2 is the attenuation coefficient of the exponentially decaying wave and this coefficient is infinite.

Inside the potential well there are two waves one forward and the second backward and both have equal amplitude but opposite sign so that at the two boundaries of the potential well the sum of the two produce nodes.

$\psi_{0\text{incident}} - \psi_{0\text{reflected}} = 0$ at the boundaries 1.60

Therefore

$\psi(z,t)_{\text{incident}} = A\,\text{Exp}[j(\omega t - k_{z1} .z)]$

$\psi(z,t)_{\text{reflected}} = -A\text{Exp}[j(\omega t + k_{z1} .z)]$ 1.61

The superposition of the two waves is:

$\psi(z,t) = A\mathrm{Exp}[j(\omega t)]\ [\ \mathrm{Exp}(- k_{z1}.z)- \mathrm{Exp}(+ k_{z1}.z)]$

$\psi(z,t) = A\mathrm{Exp}[j(\omega t)][\mathrm{Cos}(k_{z1}.z)-i\mathrm{Sin}(k_{z1}.z) - \mathrm{Cos}(k_{z1}.z)-i\mathrm{Sin}(k_{z1}.z)]$

$\psi(z,t) = -2iA\mathrm{Exp}[j(\omega t)][\mathrm{Sin}(k_{z1}.z)]$

or $\boldsymbol{\psi(z,t) = B\mathrm{Exp}[j(\omega t)]\ [\mathrm{Sin}(kz1.z)]}$ **1.62**

Eq.(1.62) is the mathematical formulation of the standing wave.

According to boundary conditions, this standing wave is suppose to have two nodes :

$\psi(0,t) = \psi(W,t) = 0$

therefore $k_{z1}.W = (2\pi/\lambda_n)W = n\,\pi$ where n = 1, 2, 3,………….

Therefore $\boldsymbol{W = n.\ \lambda n/2}$ **1.63**

This is the same restriction as was imposed in Eq.(1.51). This Eq(1.63) sheds a very important light on the quantization of energy levels permitted for an electron in an infinite potential well.

Inside the potential well there is the superposition of incident and reflected wave. There are only certain cases for which constructive interference takes place for the remaining cases there is destructive interference.

The cases for which the superposition leads to node formation at the boundaries $z = 0$ and W, only for those cases electron has an existence. For all the remaining cases electron goes out of existence by destructive interference.

Therefore we assert that in an infinite potential well , the electron can be in existence only for the energy levels:

$E = E_0 , 4 E_0 , 9 E_0 , 16 E_0$ ………

Where $E_0 = h_2/(8mW_2)$

The above result comes from Eq.(1.50) and its previous line.

An identical situation prevails in a hydrogen atom. The orbits which allow constructive interference to take place only those orbits are permitted. Remaining are forbidden. The orbital radii which support constructive interference are the radii where electron can exist as standing wave.

These are the radii exactly predicted by Bohr's Law:

Orbital Angular Momentum = Integral Multiple of $(h/(2\pi))$ as already seen in Chapter 1_part 6.

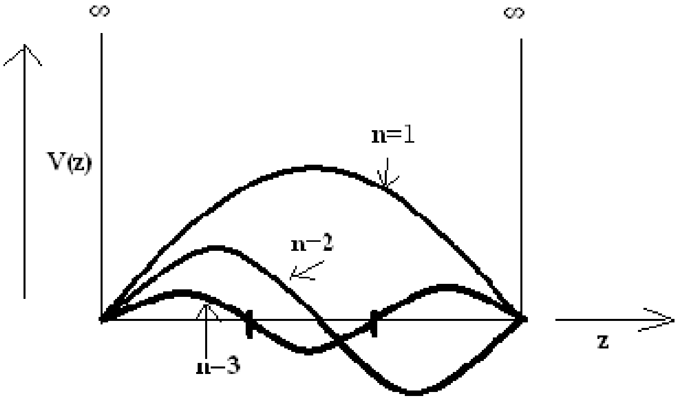


Figure 1.27. Electron's probability amplitude in an infinite potential well for the lowest three quantum states.

Fig(1.27) An electron in 3 quantum states n = 1,2,3 in an infinite potential well.

An electron in first three permissible quantum states in an infinite potential well are shown in Fig.(1.27

SSPD_Chapter1_Part 8_continued_Electron in an Finite 1-D Potential Well_Solution of Schrodinger Equation.
SSPD_Chapter 1_Part 8 is continued. We saw the solution of Schrodinger Equation for an electron trapped in an INFINITE potential well. Here we examine an electron in FINITE height potential well. The electron does not have sufficient energy to climb up the potential well still we find that it has exponentially decaying probability of finding in the potential barrier as it penetrates into the barrier.This is counter intuitive from classical point of view.

SSPD_Chapter 1_Part 8_continued_AN ELECTRON IN A FINITE HEIGHT 1-D POTENTIAL WELL.

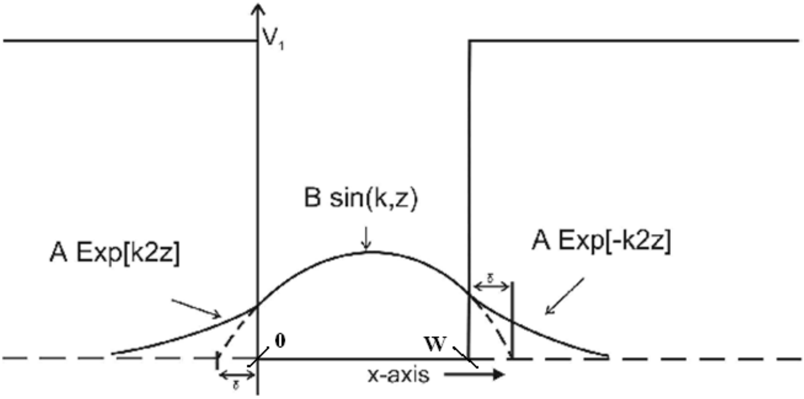


Figure 1.28 Electron in first quantum state in a finite potential well

Fig(1.28) Electron in first quantum state in a finite potential well.

Electron total energy E is less than V_1 hence classically it cannot come out of the well.

Inside the potential well for $0 < z < W$:

$\partial^2 \psi / \partial z^2 + (2mE/\hbar^2)\psi = 0$ 1.64

Outside the potential well for $z < 0$ and $W < z$:

$\partial^2 \psi / \partial z^2 + (2m(E-qV_1)/\hbar^2)\psi = 0$ 1.65

The wave vector inside the potential well is :

k_1 and $k_1^* = \pm i[\sqrt{(2mE)}]/\hbar$;

The wave vector outside the potential well is:

$k_2 = -[\sqrt{(2m(qV_1-E)}]/\hbar$;

Inside the potential well we have harmonic solution:

That is $\psi(z,t) = B\mathrm{Sin}[k_1(z+\delta)]\mathrm{Exp}[i\omega t]$

Where $k_1 = 2\pi/\lambda_1$ and by inspection $\lambda_1 = 2(W+2\delta)$ and

the solution is a standing wave with nodes at $z = -\delta$ and at $z = (W+\delta)$;

Outside the potential well we have exponentially decaying solution:

For $z < 0$, we have $\psi(z,t) = A\mathrm{Exp}[k\ 2\ z].\mathrm{Exp}[i\ \omega t]$

For $W < z$, we have $\psi(z,t) = A\mathrm{Exp}[-k\ 2\ (z-W)].\mathrm{Exp}[i\ \omega t]$ 1.66

To satisfy the boundary condition:

At $x=0$, $A= B\ \mathrm{Sin}[\pi\ \delta/(W+2\ \delta)]$

At $x=W$, $A= B\ \mathrm{Sin}[\pi\ (W+\delta)/(W+2\ \delta)]$ 1.67

Since $\pi - \pi\ (W+\delta)/(W+2\ \delta) = \pi\ \delta/(W+2\ \delta)$

Therefore the two boundary conditions are:

$A= B\mathrm{Sin}(\pi-\Theta)$;

$A = B\mathrm{Sin}(\Theta)$

where $\Theta= \pi\ (W+\delta)/(W+2\ \delta)$;

This a consistent boundary condition therefore the solution is a valid solution .

Fig(1.29) depicts an electron in second quantum state $n= 2$.

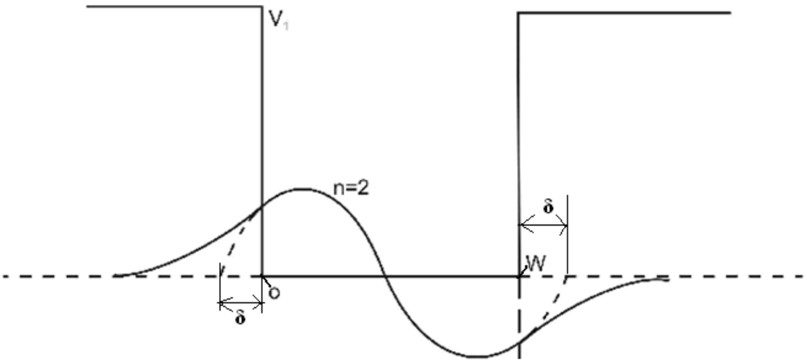


Figure 1.29. Electron in the second quantum state n=2.

Fig(1.29) Electron Matter Wave in second quantum state n = 2;

Thus here we get a counter-intuitive result. Classically electron should remain confined to the potential well because it has insufficient energy to climb over the potential barrier at the two nodes but we find that electron has exponentially decaying probability of being found inside the Potenial Barrier region as it penetrates into the barrier. As a consequence we donot find the node of the standing wave pattern at the edges of the well. Infact it has shifted inside the barrier as shown in the Figure 1.29.

SSPD_Chapter1_Part 8_continued_Electron in an 1-D Potential Well of finite height and finite width_Quantum Mechanical Tunneling

SSPD_Chapter 1_Part 8 is continued to understand the situation when we have relatively thin potential barrier and moderately high potential barrier. We saw in the last section that matter wave is capable of penetrating a potential barrier and if the barrier is thin enough then there is a possibility that matter wave may emerge out of the other side of the barrier in a much attenuated form. This phenomena is called Quantum Mechanical Tunneling. We conclude the chapter by determining the tunneling ratio.

SSPD_Chapter 1_Part 8_continued_AN ELECTRON IN A POTENTIAL WELL OF FINITE HEIGHT AND FINITE WIDTH – QUANTUM MECHANICAL TUNNELING.

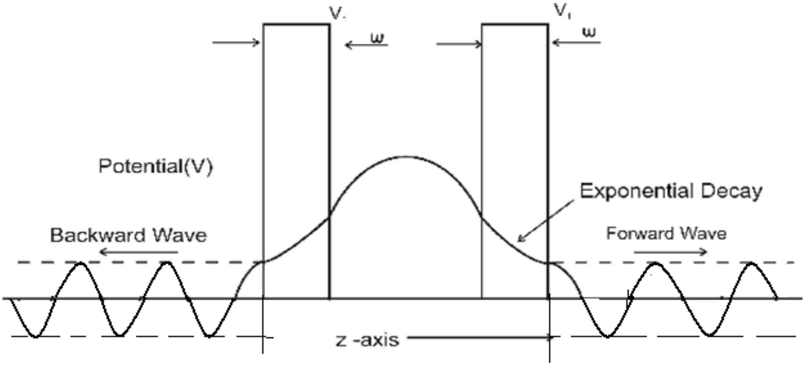


Figure 1.30. An electron in a Potential Well of finite height and finite width.

Fig(1.30) Depiction of quantum mechanical tunneling in a potential well of finite width.

Classically an electron trapped in a potential well cannot come out of it unless it has sufficient energy to cross the potential barrier confining the electron in a potential well. In photo-ionic emission as already seen and in thermo-ionic emission as we will see in the second chapter, when the photon energy imparted or thermal energy acquired exceeds the work function = $q\varphi_{surface}$ (where $\varphi_{surface}$ is the surface barrier potential) then only electron escapes into vacuum from the cathode electrode. But for an electron trapped in a potential well there is a finite probability that the electron will appear outside the well if the potential barrier is sufficiently thin. This phenomena is known as quantum mechanical tunneling.

In Fig(1.30), the phenomena of quantum mechanical tunneling is illustrated.

Within the well electron exists as a standing wave . Outside the well the electron exists as a traveling wave.

On the right of the well it exists as a forward traveling wave and on the left of the well it exists as a backward traveling wave.

For the mathematical simplicity we analyze an electron tunneling through a finite width potential barrier of finite height as depicted in Fig(1.31). The electron is travelling from left to right.

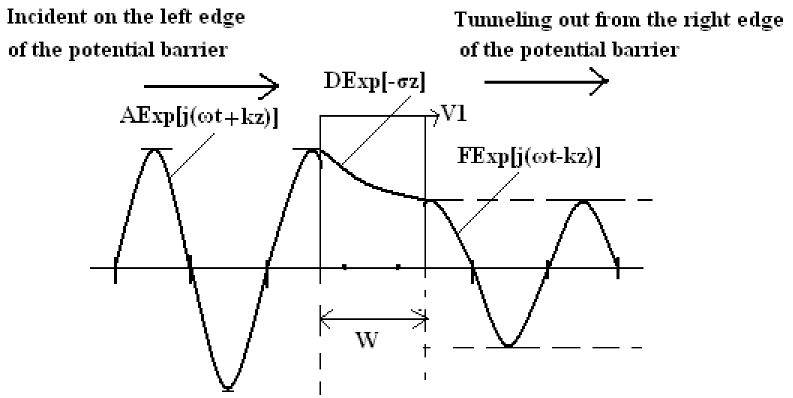

Incident on the left edge of the potential barrier
Tunneling out from the right edge of the potential barrier
DExp[-σz]
AExp[j(ωt+kz)]
V1
FExp[j(ωt-kz)]
W

Figure 1.31. Quantum Mechanical tunneling through a potential barrier of height V1 and width W.

Fig(1.31) Quantum Mechanical Tunneling of Potential Barrier of height V 1 and width W.

According to Fig(1.31) the Schrodinger Equation on the left is:

$$\partial_2 \psi / \partial z_2 + (2mE/\hbar_2)\psi = 0$$

The wave vectors are k_1 and $k_1^* = \pm i[\sqrt{(2mE)}]/\hbar$

And the solution is:$\boldsymbol{\psi I = AExp\,[i(\omega t + |k1|z)] + B.\ Exp\,[i(\omega t - |k1|z)]}$ **1.68**

Eq.(1.68) is a combination of forward and backward traveling waves.

Inside the potential barrier:

$\partial_2 \psi /\partial z_2 - \{2m(qV_1-E)/\hbar_2\}\psi = 0$

The roots of the characteristic equation are real hence the roots are the attenuation coefficients: $\sigma_1 = [\sqrt{(2m(qV_1-E)}]/\hbar$ and $\sigma_2 = -[\sqrt{(2m(qV_1-E)}]/\hbar = -\sigma_1$;

And the solution is: **ψII = [C.Exp(σ1z)+ D.Exp(-σ1z)] Exp [i(ωt)] 1.69**

Eq.(1.69) is not a traveling wave but a combination of exponentially decaying and exponentially growing function out of which exponentially growing term is technically not feasible.

On the right of the barrier:

$\partial_2 \psi /\partial z_2 + (2mE/\hbar_2)\psi = 0$

The wave vectors are again k_1 and $k_1^* = \pm i[\sqrt{(2mE)}]/\hbar$

And the solution is:**ψIII = E.Exp [i(ωt + |k1|z)] + F. Exp [i(ωt - |k1|z)] 1.70**

According to the boundary conditions: $\psi_I(x=0) = \psi_{II}(x=0)$;

$\psi_{II}(x=W) = \psi_{III}(x=W)$;

For the continuity of the wave function at the boundary we have:

$\partial\psi\ I\ /\partial z|(x=0)\ =\ \partial\psi\ II\ /\partial z|(x=0);$

$\partial\psi\ II\ /\partial z\ |(x=W)\ =\ \partial\psi\ III\ /\partial z|(x=W);$ **1.71**

Applying the above boundary conditions to the three solutions of the Schrodinger Equations we get the following:

$\psi_I\ =\ AExp[i(\omega t\ +\ |k_1|z\ )]$ – Forward traveling wave incident on the barrier from the left;

$\psi_{II}\ =\ [D.Exp(-\sigma_1 z)]Exp[i(\omega t)]$- Exponentially decaying function within the barrier;

$\psi_{III}\ =\ F.\ Exp[i(\omega t\ -\ |k_1|z\ )]$- Transmitted traveling wave on the right of the barrier after quantum mechanical tunneling;

Also $A\ =\ D$ **and** $F\ =\ D.Exp(-\sigma 1W)$ **1.72**

$|A|_2\ =$ probability density of the incident electron wave;

$|F|_2\ =$ probability density of the transmitted wave;

Therefore quantum tunneling ratio =

$T=|F|\ 2\ /|A|\ 2\ =Exp[-2W\{\sqrt{(2m\ e\ (qV\ 1\ -E))}\}/\hbar]$ **1.73**

This Eq.(1.73) establishes that if W is thin enough of the order of 10Å and (qV_1-E) of the order of 1eV (this is the order of built-in barrier potential at a pn junction then T is a finite quantity and quantum mechanical tunneling has good probability of occurring.

In Table (1.4) tunneling ratio is tabulated for different potential height and different potential widths.

Table 1.4. Quantum Mechanical Tunneling Ratio .

Width of the barrier(W) Å	**(qV_1-E) eV**	T
10	1	**3.5×10^{-5}**
5	1	**6×10^{-3}**
1	1	0.36
1	10	0.04
1	100	**3.5×10^{-5}**

Tunneling ratio 3.5×10^{-5} implies that in a potential barrier width of 10 Å and $(V_1-E)=1V$ out of 10^5 incident electrons 3.5 electrons will tunnel through. If the potential barrier width is reduced to 1 Å and barrier height is $(V_1-E)=1V$ then almost 36% of the incident electrons will tunnel through.

Tunnel Diode, Backward Diode, Zener Diode as well as high field emission utilizes quantum mechanical tunneling phenomena. In 1957 , Esaki **[Appendix XXXV]** invented the first quantum device namely Tunnel Diode. At University of Cambridge, Brian Josephson **[Appendix XXXVI]** adapted Tunnel Diode to superconducting materials giving rise to Josephson effect.For the invention of Tunnel Diode Brian Josephson (England), Leo Esaki (Japan) and Evar Gebar (USA) were awarded the Nobel Prize in Physics in the year 1973. In honour of Leo Esaki, Tunnel Diode was named as Esaki Diode.

SSPD_Chapter1_Part 8_conclusion_Electron travelling across a step potential

SSPD_Chapter 1_Part 8 concludes that electron behaves exactly the same fashion as light does. Light at the interface of two different optical media suffers a reflection while being transmitted across a step index. In exactly an analogous manner electron matter wave suffers a reflection while being transmitted across a voltage step. Quantum behaviour of electron becomes more pronounced as we move to nano scale of miniturization.

SSPD_Chapter 1_Part 8_conclusion_ ELECTRON TRAVELING ACROSS A STEP POTENTIAL WITH TOTAL ENERGY GREATER THAN THE POTENTIAL BARRIER OFFERED BY THE STEP POTENTIAL.

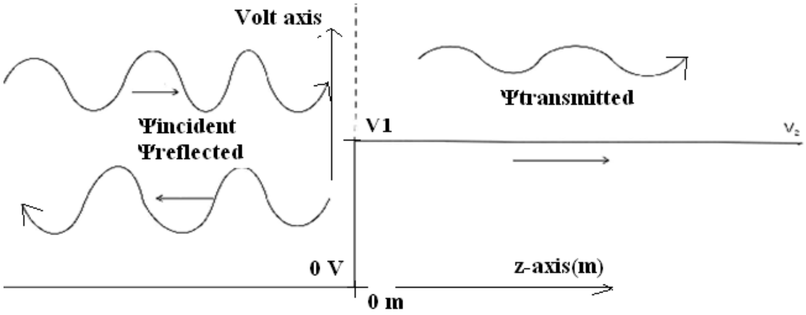


Figure 1.32. Electron travelling from left across a potential barrier of height V1. Electron has energy larger than that needed for crossing the potential barrier.

Fig.(1.32) Electron traveling across a step potential with energy greater than that needed for crossing the barrier potential.

In Figure (1.32) electron is traveling across a step potential with sufficient energy to cross the barrier. The electron matter wave experiences a partial reflection at the step potential just as light experiences a reflection when passing across step change in refractive index.

The linear momentum also changes while moving from one region of potential energy to another region of potential energy,

Thus **$p_1^2 /(2m_e) = (E)$ and wave vector $k_1 = p_1/\hbar$;**

And $p_2^2 /(2m_e) = (E - qV_1)$ and wave vector $k_2 = p_2/\hbar$; 1.74

The incident wave: $\psi_I = A\mathrm{Exp}[i(\omega t + |k_1|z)]$

The reflected wave: $\psi_R = B\mathrm{Exp}[i(\omega t - |k_1|z)]$

The transmitted wave: $\psi_T = C\mathrm{Exp}[i(\omega t - |k_2|z)]$ 1.75

In 1961 , Texas Instrument brought the first IC Chip where the feature size was several microns and which could be analyzed in the classical manner. In microchips the classical properties of the electrons were manifested.

By the turn of century PENTIUM IV INTEL Chip **[Appendix XXXVII]** has come in the market where the feature size has come below 100nm. These nanochips manifest the quantum mechanical properties of the electrons and are more appropriately called Quantum Chips.

By the use of VLSI/ULSI technology many quantum devices are coming in the market.

SSPD_Chapter 1_Part8_AppendixXXXIV_Travelling Waves

Appendix XXXIV refers to SSPD_Chapter_Part8 It gives a simple quantitative method of arriving at the solution of Wave Equations of Electro-Magnetic Field Theory. We have a similar methodology of arriving at the solution of Schrodinger Equation.

Appendix XXXIV

Solving the wave equation for plane wave

Maxwell’s four differential equations are:

$$\nabla . D = \rho$$

$$\nabla . B = 0$$

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

$$\nabla \times H = J_c + \frac{\partial D}{\partial t}$$

In free space propogation: ρ (charge density) $=0$ and

$$J_c = 0.$$

Therefore the four differential equations become:

$$\nabla . D = 0$$

$$\nabla . B = 0$$

$$\nabla \times E = -\frac{\partial B}{\partial t}$$

$$\nabla \times H = \frac{\partial D}{\partial t}$$

From these four equations we obtain:

$$\nabla \times \nabla \times E = \nabla(\nabla . E) - \nabla^2 E = -\nabla^2 E = -\mu€\frac{\partial^2 E}{\partial t^2}$$

$$\nabla \times \nabla \times H = \nabla(\nabla . H) - \nabla^2 H = -\nabla^2 H = -\mu€\frac{\partial^2 H}{\partial t^2}$$

But in free space both divergences are zero therefore the above 2 equations become:

$$\nabla^2 E = \mu€\frac{\partial^2 E}{\partial t^2}$$

$$\nabla^2 H = \mu \text{€} \frac{\partial^2 H}{\partial t^2}$$

In free space $\text{€} = \text{€}_0$ and $\mu = \mu_0$

That is free space permittivity = absolute permittivity .

And free space permeability = absolute permeability .

Also c (velocity of light in free space) =

$$\frac{1}{\sqrt{\mu_0 \text{€}_0}}$$

Therefore the above two equations become:

$$\nabla^2 E = \frac{1}{c^2} \frac{\partial^2 E}{\partial t^2}$$

$$\nabla^2 H = \frac{1}{c^2} \frac{\partial^2 H}{\partial t^2}$$

We apply separation of variables to solve these two partial differential equations:

Assume E(r,t) = E1(r)E2(t)

H(r,t) = H1(r)H2(t)

In general by assuming the time dependent portion to be of harmonic nature:

That is E2(t) = E20Exp(jωt) and H2(t) = H20Exp(jωt)

The partial derivative

$$\frac{\partial}{\partial t} = j\omega$$

Hence the two equations become:

$$\nabla^2 E1 = \frac{-\omega^2}{c^2}\frac{E1}{1}$$

$$\nabla^2 H1 = \frac{-\omega^2}{c^2}\frac{H1}{1}$$

Now the partial differential equations have reduced to ordinary linear differential equations known as wave equations of 2nd order and they can be solved applying operator theory.

We also assume that it is a plane wave travelling along z axis hence

$$\nabla^2 = \frac{\partial^2}{\partial z^2}$$

Under such circumstances as we will see in Electro-Magnetic Field Theory, only x-component of E1 and y-component of H1 will remain. Hence

$$\nabla^2 E1x = \frac{-\omega^2}{c^2}\frac{E1x}{1}$$

$$\nabla^2 H1y = \frac{-\omega^2}{c^2}\frac{H1y}{1}$$

Applying Theory of Operator we get two roots in each case namely:

Two roots are $=$

$$\pm j\,\frac{\omega}{c} = \pm j\,\frac{2\pi\nu}{c}$$

$=$

$$\pm j\,\frac{2\pi}{\lambda}$$

$=$

$$\pm jk \leftrightarrow wave\ vectors$$

Hence

$$E1_x = AExp(+jkz) + BExp(-jkz)$$

$$H1_y = CExp(+jkz) + DExp(-jkz)$$

So the complete solution for Electric Field is:

$$E(z,t) = E1_x(z)E2(t)$$

$$E(z,t) = AExp[j(\omega t + kz)] + BExp[j(\omega t - kz)]$$

$$H(z,t) = CExp[j(\omega t + kz)] + DExp[j(\omega t - kz)]$$

1st Term is backward travelling wave and 2nd term is forward travelling wave.

SSPD_Chapter 1_Part8_AppendixXXXV_Schrodinger Eqaution and its solution
SSPD_Chapter 1_Part 8_Appendix XXXV gives the setting up of Schrodinger Equation from the four postulates of Quantum Mechanics and method of solving this equation.

APPENDIX XXXV _Schrodinger Equation and its solution.

To fully understand Schrodinger Equation we must understand the concept of operator associated with any physical observable (dynamical variable) and its canonical conjugate.

Translational momentum p is associated with the spatial coordinates: x, y and z;

Rotational momentum J is associated with the curvilinear coordinate θ;

Energy E is associated with temporal element time t.

We give the following excerpt from “ Basic Quantum Mechanics” Robert L. White, Publisher McGraw Hill, 1966.

The postulates of Quantum Mechanics:

POSTULATE 1. With every physical observable ‘a’, we may associate a mathematical operator A from whose properties we may deduce the possible results of measurements of the physical observable ‘a’. Measurement of a physical observable is analogous to taking a measurement. The result of measurement is ‘a_n’.

$$Au_n = a_n u_n$$

XXXV.1

u_n is a solution of the above equation. The above equation has a set of solutions which meet certain boundary conditions. This set of solutions are known as Eigen Functions.

In the above equation, operator A operates on the eigen function ‘u_n’ and generates an eigen value ‘a_n’ . If the eigen functions exist for discrete values of ‘a_n’ then we say we have DISCRETE eigenvalue spectrum.

If the eigen functions exist for continuous values of ‘a_n’ then we say we have CONTINUOUS eigenvalue spectrum.

POSTULATE 2. The only possible result of a measurement of the physical observable ‘a’ is one of the eigen values of operator A.

POSTULATE 3. For every dynamical system there exists a state function ψ which contains all the information that is known about the system. Ψ is the probability amplitude and |ψ|2 is the probability density of obtaining a measurement ‘a_n’ while carrying out a series of measurements on observable ‘a’.

POSTULATE 4. If we know ψ for any system at any particular time , then the evolution of ψ at all subsequent times is determined by:

$$-\frac{\hbar}{i}\frac{\partial\psi}{\partial t} = H_{op}\psi$$

Where H_{op} is the operator associated with the Hamiltonian of the system. Postulate 4 directly leads to Schrodinger Equation.

Operators of Quantum Mechanics are generally not commutative.

Commutator Bracket is [A,B] = (AB-BA) = -[B,A]

[A,A] = 0

[A,K] = 0 where K is a constant.

[A+B,C] = [A,C] + [B,C]

[A,(BC)] = [A,B]C + B[A,C];

What the above properties imply is the following:

1. Operator associated with any generalized coordinate(physical observable) commutes with operator associated with any other generalized coordinate.

$[Q_j , Q_k] = 0;$

1. Likewise , the operators associated with two different generalized momenta always commute namely $[P_j , P_k]$;
2. But the operator associated with a given generalized coordinate and the operator associated with its own conjugate momentum do not commute namely $[Q_j , P_j] = i\hbar$ whereas $[Q_j , P_k] = 0$ as long as $j \neq k$.Not only do they fail to commute, but their commutator bracket is equal to a very specific quantity, $i\hbar$, involving Planck's constant. If we made a numerically different assumption for the commutator bracket then we would predict a quantitatively incorrect optical emission spectrum for the hydrogen atom or a black body radiation law at variance with experimental fact or we would obtain a wrong infrared spectrum of solids.
3. Therefore $[Q_j , P_j] = i\hbar$ is the basic assumption of Quantum Mechanics.

Schrodinger in his classic formulation proposed a set of operators for quantum mechanical dynamical variables and their canonical conjugates.

Dynamical Variable	Associated Operator
x →	x (multiply by x)
p_x →	$\frac{\hbar}{i}\frac{\partial}{\partial x}$
y →	y (multiply by y)
p_y →	$\frac{\hbar}{i}\frac{\partial}{\partial y}$
z →	z (multiply by z)
p_z →	$\frac{\hbar}{i}\frac{\partial}{\partial z}$
p →	$\frac{\hbar}{i}\nabla$
p^2 →	$-\hbar^2\nabla^2$
t →	$i\hbar\frac{\partial}{\partial t}$

We will consider the Hamiltonian Operator for conservative system:

H = T + qV where T is the Kinetic Energy and qV is the potential energy.

In translational motion:

$$H = T + qV = \frac{p^2}{2m} + qV$$

$$\boldsymbol{p}^2 = -\hbar^2 \boldsymbol{\nabla}^2$$

Therefore

$$H_{op} = -\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right) + qV(r)$$

We let this Operator operate on a probability amplitude function ψ(r,t).

Therefore:

$$H_{op}\psi = -\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2} + \frac{\partial^2\psi}{\partial y^2} + \frac{\partial^2\psi}{\partial z^2}\right) + qV(r)\psi = E\psi = i\hbar\frac{\partial\psi}{\partial t}$$

XXXV.2

Eq XXXV.2 is defined as Schrodinger Equation.

Just as we used the separation of variables in solving the Wave Equation in Appendix XXXIV we will use separation of variables in this equation also. In addition we will consider this as 1-D problem with V(x) = 0.

Hence Schrodinger Equation reduces to:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2}\right)=i\hbar\frac{\partial\psi}{\partial t}$$

Let ψ(x,t) = ψ1(x).ψ2(t)

We assume ψ2(t) = Exp(-iωt) hence

$$i\hbar\frac{\partial\psi}{\partial t}= i\hbar\psi 1(x)(-i\omega\psi 2(t))$$

Therefore Right hand side of the Schrodinger equation is :

RHS = $+\hbar$ω

$$\psi 1(x)\psi 2(t)$$

Therefore Schrodinger Equation reduces to:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2}\right)=\hbar\omega\ \psi 1(x)\psi 2(t)$$

This simplifies to:

$$-\frac{\hbar^2}{2m}\ \psi 2\left(\frac{\partial^2\psi 1}{\partial x^2}\right)=\hbar\omega\ \psi 1(x)\psi 2(t)$$

Ψ2(t) cancels out from both sides hence we have an ordinary linear differential equation namely:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi 1}{\partial x^2}\right) = \hbar\omega\,\psi 1(x)$$

$$\left(\frac{\partial^2\psi 1}{\partial x^2}\right) = -\frac{2m\omega}{\hbar}\,\psi 1(x)$$

This is a second order ordinary differential equation with two roots namely ±i√(

$$\frac{2m\omega}{\hbar})$$

√(

$$\frac{2m\omega}{\hbar}) = \sqrt{\left[(2m)(2\pi\nu)\left(\frac{2\pi}{h}\right)\right]}$$

=

$$\sqrt{(\frac{2\pi}{h})^2 \times 2m \times h\nu}$$

=

$$\sqrt{(\frac{2\pi}{h})^2 \times 2m \times E}$$

But E =

$$\frac{p^2}{2m}$$

Therefore

$$\sqrt{(\frac{2\pi}{h})^2 \times 2m \times E}$$

=

$$\sqrt{(\frac{2\pi}{h})^2 \times p^2}$$

=

$$\frac{2\pi}{\lambda}$$

= k* = wave vector.

Hence ±√(

$$\frac{2m\omega}{\hbar})$$

= ±k*

Therefore the solution is

Ψ(x,t) = ψ1(x)ψ2(t)

And ψ1(x) = AExp(+ik*x) + Bexp(-ik*x)

Therefore total solution is : Ψ(x,t) = ψ1(x)ψ2(t)

Or Ψ(x,t) = [AExp(+ik*x)+ Bexp(-ik*x)]Exp(-iωt)

Or Ψ(x,t) = AExp[-i(ωt-k*x)] + Bexp[-i(ωt +k*x)]→ These are two travelling waves: 1st is moving forward and second is moving backward.

If V(x) = V_1 = constant then the Schrodinger equation will be :

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2}\right)+qV_1\psi = i\hbar\frac{\partial\psi}{\partial t}$$

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2}\right)+qV_1\psi = i\hbar\frac{\partial\psi}{\partial t}$$

Again we apply the separation of variables method:

Let ψ(x,t) = ψ1(x).ψ2(t)

We assume ψ2(t) = Exp(-iωt) hence

$$i\hbar\frac{\partial\psi}{\partial t} = \ i\hbar\psi1(x)(-i\omega\psi2(t))$$

Therefore Right hand side of the Schrodinger equation is :

RHS $= \ +\hbar\omega$

$$\psi1(x)\psi2(t)$$

Therefore:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi}{\partial x^2}\right)+qV_1\psi =$$

$+\hbar\omega$

$$\psi1(x)\psi2(t)$$

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi1}{\partial x^2}\right)+qV_1\psi1 =$$

$+\hbar\omega$

$$\psi1(x)$$

Rearranging the terms:

$$-\frac{\hbar^2}{2m}\left(\frac{\partial^2\psi 1}{\partial x^2}\right)=$$

$+\hbar\omega$

$$\psi 1(x)-qV_1\psi 1(x)$$

$=\ (\hbar\omega$

$$-qV_1)\psi 1(x)$$

$$\left(\frac{\partial^2\psi 1}{\partial x^2}\right)=-\frac{2m}{\hbar^2}$$

$(\hbar\omega$

$$-qV_1)\psi 1(x)$$

$=$

$$-\frac{2m}{\hbar^2}$$

$(E$

$$-qV_1)\psi 1(x)$$

This also has two roots : ±i

$$\sqrt{\frac{2m}{\hbar^2}(E - qV_1)}$$

= ±ik**

Therefore the solution is: Ψ(x,t) = AExp[-i(ωt-k**x)] + Bexp[-i(ωt+k**x)]

If $E > qV_1$, we have imaginay roots of the ordinary linear differential equation and we get traveling waves as the solution.

But if $E < qV_1$, we have real roots of the ordinary linear differential equation and we get exponentially decaying or exponentially growing solutions.

SSPD_Chapter 1_Part 9_PAULI-EXCLUSION PRINCIPLE

Part 9 of Chapter 1 describes the four quantum numbers and Pauli-Exclusion Principle.

SSPD_Chapter 1_Part 9_ THE FOUR QUANTUM NUMBERS OF ATOMIC ELECTRONS AND PAULI-EXCLUSION PRINCIPLE.

Till now we have solved the Schrodinger Equation for 1-D Potential Well. If the same equation is solved for a real life 3-DPotential Well associated with a Hydrogen Atom then we get the following equation:

$$\nabla^2\psi - \frac{2m_e(E-V(r))}{\hbar^2} = 0$$

1.74

The solution of this Equation gives the full quantum mechanical description of electron bound to the host atom in this case hydrogen atom.

The potential well of a hydrogen atom or any real atom has a spherical symmetry hence we use spherical coordinates r, Θ, φ, while solving Eq. (1.74)

By taking the nonrelativistic Hamiltonian, we get three quantum numbers n, l, and m.

n is the principal quantum number or the total quantum number and specifies the K,L,M,N…… Shells of the electron orbits It gives the quantized values of energies associated with these principal Shells.

In Hydrogen Atom , the principal energy levels are:

$$E_n = \frac{-13.6eV}{n^2}$$

Table 1.5. Energy Levels of principal quantum numbers.

n	**E n**
1	13.6 eV
2	3.4eV
3	1.5eV
4	-0.85eV

The energies also known as eigen energies associated with the principal quantum number are negative since we have taken the energy of an electron removed far away from the atom as zero and we are considering the energy of bound electrons. Therefore the eigen energies are negative.

There are infinite such bound states becoming very closely spaced for large values of n.

The second quantum number ‘l’ is azimuthial or angular momentum quantum number. It gives the quantization of orbital angular momentum L as well as the shape of the orbital sub-shell belonging a given shell n.

$L = I\omega =$ Moment of inertia of the orbiting electron × angular orbital velocity;

For a given n, $l = 0, 1\ 2 \ldots\ldots.(n-1)$

i.e. $n = 1$, $l = 0$; only one value of l;

$n = 2$, $l = 0, 1$;

$n = 3$; $l = 0, 1, 2$;

The magnitude of orbital angular momentum $|L| = [\sqrt{\{l\ (l + 1)\}}] \times \hbar$

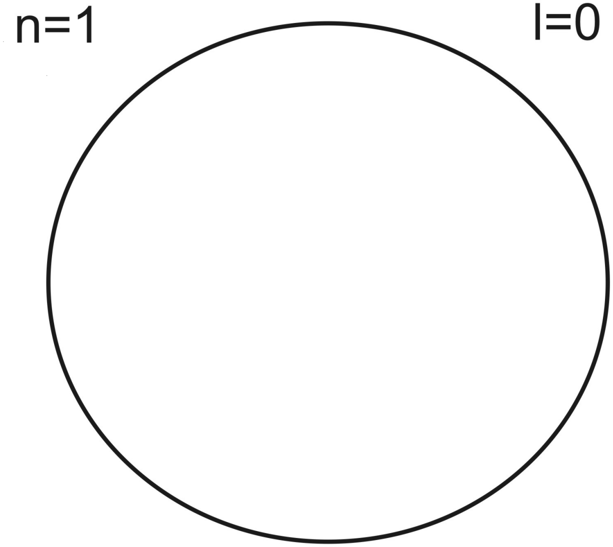
And $|L|_{max} = (l + \Delta) \times \hbar$;

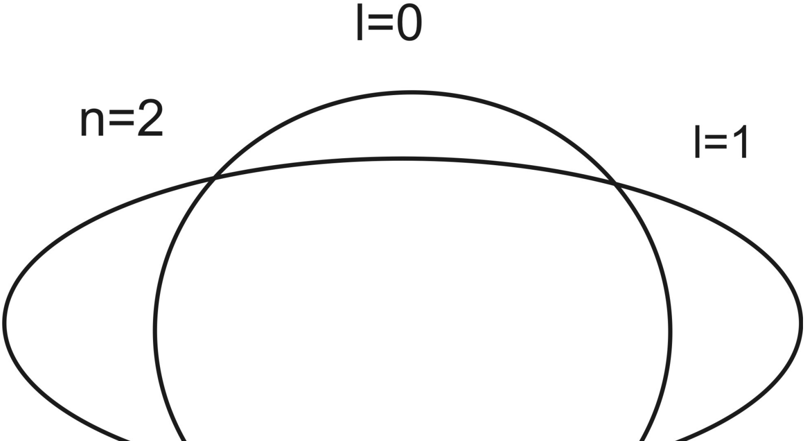
Therefore $l = 0$ implies zero orbital angular momentum and a perfectly spherical sub-shell belonging to a given shell n.

$l = 0$ constrains the orbital sub-shell to be have a spherical symmetry and therefore all possible orbital angular momentum vectors cancel one another to give a resultant $|L| = 0$.

$l = 1$ implies $|L| = [\sqrt{2}] \times \hbar$. This implies a resultant orbital angular momentum and hence an ellipsoidal sub-shell.

$l = 2$ implies $|L| = [\sqrt{2(2+1)}] \times \hbar = [\sqrt{6}] \times \hbar$. This implies a higher oblateness of the ellipsoidal sub-shell.


n=1
l=0


l=0
n=2
l=1

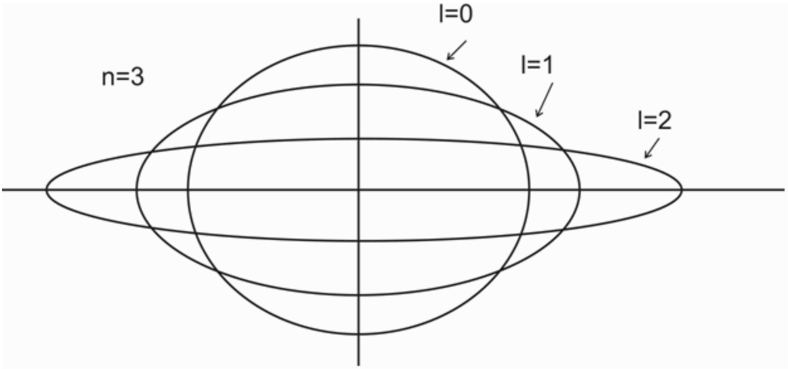


Figure 1.33. The possible permitted subshells of electrons belonging to a host atom.

From nonrelativistic treatment there is n-fold degeneracy for every value of principal quantum number ‘n’. This means the electrons in given shell ‘n’ have different values of ‘l’ ranging from 0 to n-1 and all these have same energy level corresponding to ‘n’. This is termed as n-fold degeneracy.

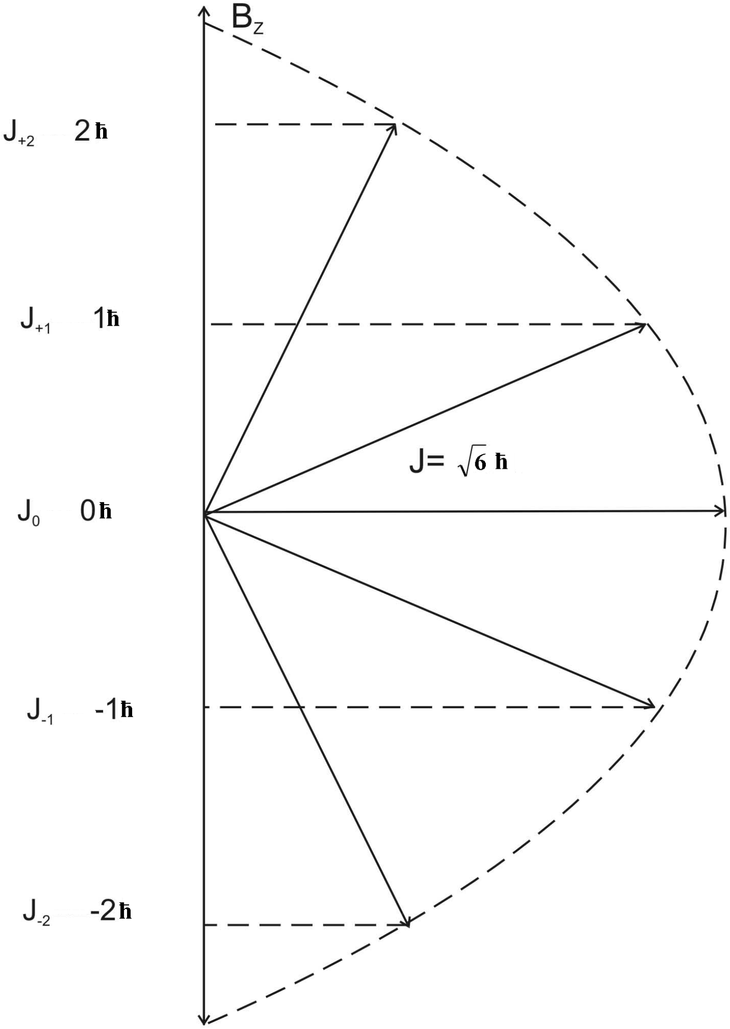
For n=1, l = 0 , there is no degeneracy

For n=2, l = 0, 1 and there is two fold degeneracy;

For n = 3, l = 0, 1 , 2 and there is three fold degeneracy;

Next we have magnetic quantum number ‘m’. This implies the spatial quantization of orbital angular momentum. For a given value of l we have a given value of Orbital Angular Momentum.

Say we take l = 2. This has L= [√6] × ħ. Now when a magnetic field is applied then L vector can be permitted to assume only those spatial directions in which the z axis projections of L i.e. Lz = 0ħ, ±1ħ, ±2ħ, as shown in Figure(1.34).



In the presence of a magnetic field, the spatial quantization of L=[$\sqrt{6}$]X ħ]

Figure 1.34. In the presence of a magnetic field, the spatial quantization of L = [√6] × ħ;

Therefore l will range from 0 to (n-1) and each value of l between 0 and n-1 has 2l + 1 eigen states.

In Fig.(1.34) L corresponding to l = 2 can be parallel or anti-parallel giving a projection of ±2ħ on z axis. L can be inclined to z-axis to give a projection of ±ħ or L can be transverse to z-axis giving a projection of 0ħ.

L corresponding to l = 0 is transverse to the z-axis. L corresponding to l = 1 is obliquely aligned with respect z-axis so as to give a projection of ±1ħ on the z-axis and L is also transverse to z-axis.

Therefore a given value of n has altogether Σ(2 l + 1) degeneracy with l ranging from 0 to (n-1) . The total number of eigen states associated with principal quantum number n is n2. These eigen state energies are not explicitly determined by l or m. Hence these n2 eigen states have the same energy level –(13.6/ n2). Therefore a given principal quantum number n has n2 -fold degeneracy. Energy states for a hydrogenic atom is given Fig.(1.35).

As seen in Figure 1.35. for n =1, l = 0; This has 1-fold degeneracy.

For n =2 , l = 0 and 1; Here there is 1-fold degeneracy for l = 0 and 3 fold degeneracy for l =

1;

For n= 3, l = 0,1 and 2; Here there is 1-fold degeneracy for l = 0 and 3 fold degeneracy for l = 1 and 5-fold degeneracy for l =2;

So we see that if in an atom the electronic shells are occupied upto n=3 then it has (1+3+5) =9 electrons. For every quantum state characterized by n, l and m there is further possibility of ±(1/2)ħ spin angular momentum.

for n =1, l = 0; This has 1-fold degeneracy. But l =0 has 2-fold degeneracy due to spin quantum number namely ±(1/2)ħ.

For n =1 there will be two electrons permitted.

For n= 2 there will be eight electrons permitted.

For n= 3 there will be eighteen electrons permitted.

And for n in general there will be 2n2 electrons permitted. This sets the rule by which the elements are built up in a Periodic Table.

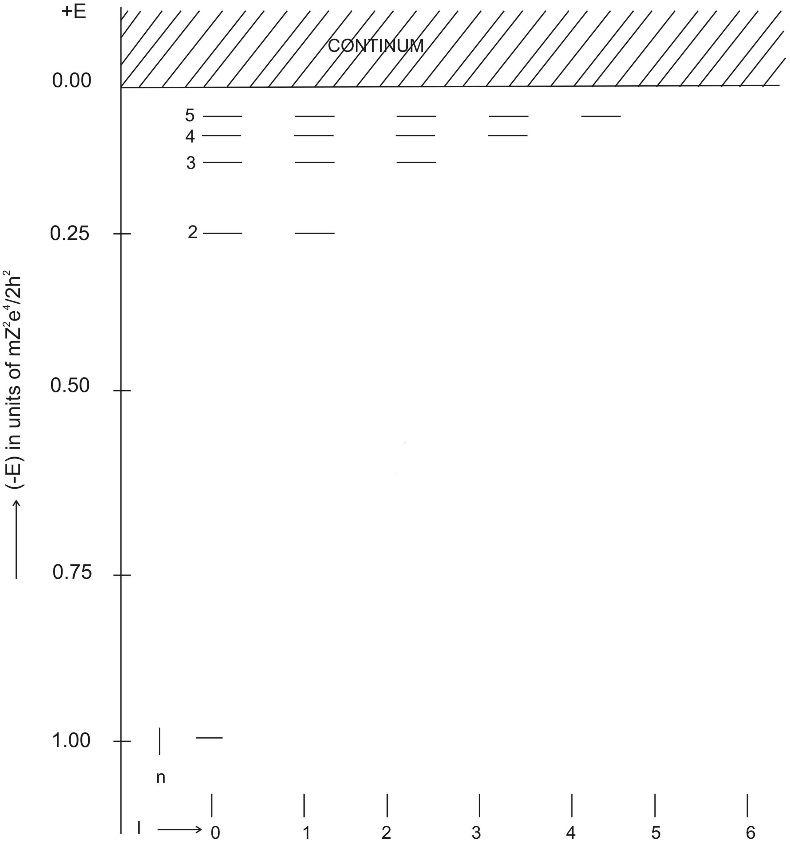


Figure 1.35. Energy Level of Hydrogenic Atom from nonrelativistic Hamiltonian.

From the discussion above and from the discussion which will follow in the sequel part, we can conclude Pauli-Exclusion Principle namely:

No two electrons can have all the four quantum numbers to be the same.

SSPD_Chapter 1_Part 9_Conclusion_Removal of degeneracy
Concluding Part 9 describes the wave nature of distinguishable and undistinguishable particles and the resulting Fermi-Dirac distribution and Bose-Einstein distribution respectively. This part describes the multiple fine structure in spectrum is caused due to the removal of degeneracy.

SSPD_Chapter1_Part9_Continued_RELATIVISTIC EFFECTS IN THE HYDROGEN ATOM.

Relativistic Hamiltonian can be fully solved for the Hydrogen atom. Here only the end results are enumerated.

As already seen azimuthial quantum number decides the shape of the subshell corresponding to different l. Now we know from Kepler's second law **[Appendix XXXIII]** that a body in elliptical orbit sweeps equal area in equal time. Hence in spherical orbit electron moves with uniform linear velocity all the time but in ellipsoidal orbit apogee velocity is the slowest and perigee velocity is the fastest. If relativistic considerations are made then n fold degeneracy is removed among the different values of l but same n.

Relativistic considerations lead to spin angular momentum of electron. The spin quantum number is $s = \pm(1/2)\hbar$.

Table 1.6. Quantum Numbers of an Atomic Electron.

Name	Symbol	Possible Values	Quantity determined
Principal	n	1, 2, 3, ….	Electron Energy
Azimuthial or Orbital	l	0,1, 2, 3, 4, ….(n 1);	
Magnetic	m_l	-l, -(l -1),… 0….(l -1), l	Spatial direction of L in presence of Magnetic field
Spin Angular Momentum	s	-1/2, +1/2	Electron Spin Angular Momentum

From the point of view of quantum numbers, following are the permissible shells and subshells:

Table 1.7. Permissible Shells and Subshells

n	l		m		s		subshell	Shell
	s		K					
	L							
	M							

Thus we see that the permissible electrons are subdivided in Major Shells:

K, L, M, N, O, P… where n = 1, 2, 3, 4, 5, 6 …… respectively.

Every Shell has Subshells.

n = 1, K has only s subshell having two electrons differentiated by spin quantum number $\pm 1/2\hbar$.

n= 2 , L shell has two subshells, s and p subshell. s shell can accommodate at most 2 electrons .

Within L shell, p subshell can accommodate 6 electrons.

Therefore L shell can accommodate 8 electrons altogether.

n=3, M shell has 3 subshells, s. p and d subshells.

s shell can accommodate 2 elecrons,

p shell can accommodate 6 electrons.

d subshell can accommodate at most 10 electrons.

Therefore M shell can accommodate 18 electrons altogether.

Accordingly nth shell can accommodate $2n^2$ electrons at most and these $2n^2$ electrons are subdivided in s, p, d, f, subshells.

As seen from Table(1.7), no two electrons have the same set of four quantum numbers. This is known as Pauli-Exclusion Principle **[Appendix XXXIV]** .

SSPD_Chapter 1_Part 9_conclusions_1.8.2_FAR REACHING IMPLICATIONS OF THE REMOVAL OF DEGENERACY.

In Chapter 1_Part 7_Sec(1.6.3.) we had derived the formula for determining the energy of the electron in shell having n principal quantum number. According to Eq.(1.35):

$$E_n = -13.6\text{eV}/n^2$$

According to this formula $2n^2$ electrons at n quantum number will all have the same energy. This implies that all the n^2 electrons are at the same energy level. If this was to be true then we would say that all n^2 electronsare in a degeneratestatebut in fact this is not true. If relativistic considerations and spin considerations are applied then the degeneracy is removed. This has far reaching consequences in terms of the type of particles we are considering.

Broadly speaking there are two broad categories of particles:

BOSONS and FERMIONS.

Bosons are virtual particles and any number of Bosons can be accommodated at the same energy E at the same time t. One example is Photons. These have a spin angular momentum of $n\hbar$ where $n = 0$, 1, 2,…..Photons have spin angular momentum $0\hbar$.These are identical particles which cannot be distinguished. They have no distinguishing characteristics. Simultaneously two or more such particles can have the same spatial coordinates as well as same momentum (x,y,z, p_x, p_y, p_z). This implies that in phase space (a combination of spatial space and momentum space), the elemental phase space ($\Delta x\ \Delta y\Delta z\ \Delta p_x\Delta p_y\Delta p_z$) can accommodate two or more particles, in the case photons. This is another way of saying that Heisenberg Principle of Uncertainty is not applicable to Bosons. These follow Bose-Einstein Statistics. A Helium Atom with 2 protons, 2 neutrons and 2 electrons have $0\hbar$ spin angular momentum hence they behave like Bosons and at liquid He temperature that is at 4K they become superfluid as we will see in later section.

The credit for the discovery of this particle goes to Satyendra Nath Bose. The story of this discovery is as follows.

At the turn of 19-20th century Black Body Radiation was an enigma. There was no correct theoretical formulation of the same. Rayleigh- Jean formulation ended in Ultra Violet Catastrophe and if the shorter wavelength could be explained then longer wavelengths could not be explained.

In 1910, Max Planck came up with an empirical formulation:

$$\langle E\rangle = \hbar\omega/[\mathrm{Exp}(\hbar\omega/(kT)) - 1] \quad 1.75$$

This empirical formulation had its roots in the quantum nature of electro-magnetic energy. Once the photon or energy packet nature of electromagnetic waves was accepted, the above equation could be theoretically derived and that is precisely what S.N Bose did.

This resulted in a completely new behavior and a new probability distribution namely:

$$F(E) = 1/[\mathrm{Exp}(E/(kT)) - 1] \quad 1.76$$

Here F(E) is the Probability of Occupancy at energy level E.

This formulation was derived for indistinguishable photon particles whose number is not conserved and submitted to a British Journal by S.N.Bose in 1924. But it was rejected. Next he sent it to Einstein who translated it and submitted it to a German Journal and it was published. Subsequently he extended this to material particles such as $_2He_4$ where total spin angular momentum is zero and hence behaves as Boson. Therefore this new probability distribution is called Bose-Einstein Statistics. Einstein predicted the fifth state of matter Bose-Einstein Condensate which was subsequently achieved in 90's.

Photons are indistinguishable whereas electrons are distinguishable due to the four quantum numbers.

What does indistinguishability and distinguishability imply.

1.8.2.1. DISTINGUISHABILITY AND INDISTINGUISHABILITY.

Let us consider a system of two particles ,1 and 2. One of them is in state a and the other in state b. If the particles are distinguishable then the pair in states a and b and the pair in state b and a are also distinguishable. The two possibilities are:

$$\psi_I = \psi_a(1).\psi_b(2) \text{ and } \psi_{II} = \psi_a(2).\psi_b(1) \quad 1.77$$

For distinguishable particles, we cannot say which set is valid. Both may be equally valid. Therefore we have two kinds of linear combination. If the wave function remains unaffected by exchange between the two states then we have symmetric function and is given as follows:

$$\psi B = (1/\sqrt{2})[\psi a(1).\ \psi b(2) + \psi a(2).\ \psi b(1)]$$
1.78

This is the wave function describing Bosons which do not obey Pauli Exclusion Principle and any number of Bosons can stay in the same quantum state.

Bosons have spin angular momentum = $n\hbar$ where $n=0,1,2,3....$

If the wave function changes sign on interchange of state then we have anti symmetric wave function and is given as follows:

$$\psi F = (1/\sqrt{2})[\psi a(1).\ \psi b(2) - \psi a(2).\ \psi b(1)]$$
1.79

This is the wave function describing Fermions which obey Pauli Exclusion Principle and no two Fermions can stay in the same quantum state.

Fermions have spin angular momentum = $(1/2 + n)\ \hbar$ where $n=0,1,2,3...$

$(1/\sqrt{2})$ has been used for normalization so that the total probability of finding the particle over the volume concerned is Unity.

Now what is the possibility that both particles are in the same quantum state (a).

For distinguishable particles, this new state of overlap of wave function is:

ψ M = ψ a (2). ψ a (1) 1.80

For bosons the wave function is:

$\psi_B = (1/\sqrt{2})[\psi_a (1). \psi_a(2) + \psi_a (2). \psi_a(1)]$

$= (1/\sqrt{2})[2\psi_a (1). \psi_a(2)] = \sqrt{2}\ \psi_a (1). \psi_a(2)$

Probability density of Bosons in same state:

$\psi_B. \psi_B^* = \sqrt{2}\ \psi_a (1). \psi_a(2). \sqrt{2}\ \psi_a^*(1).\psi_a^*(2)$

$\psi_B. \psi_B^* = 2\ \psi_M. \psi_M^*$

Thus the probability that two bosons are in the same state is twice the probability that two distinguishable particles will be in the same state.

For fermions the wave function is:

$\psi_F = (1/\sqrt{2})[\psi_a (1). \psi_a(2) - \psi_a (2). \psi_a(1)] = 0$

Therefore fermions can never be in the same quantum state. Hence fermions always follow Pauli’s Exclusion Principle.

1.8.2.2. BOSE-EINSTEIN CONDENSATE- FIFTH STATE OF MATTER.

In 1924, Einstein had visualized that in a cold enough gas, which does not become liquid or solid, loss of momentum will lead to the spread of the wave function on the container's spatial dimension scale from Heisenberg Uncertainty Principle. In case of bosons as in case of $_{2}He^{4}$ the wave functions will overlap leading to all the atoms being confined to the same quantum state leading to super atom. This is called Bose-Einstein Condensate.

In 1995, Eric Cornell, Carl Wieman and their coworkers **[Appendix XXXV]** at Colorado achieved such a condensate by cooling rubidium gas to 10^{-7}K. About 2000 rubidium atoms formed a single entity namely super atom 10 microns long and it lasted 10 seconds. In 1997 the Nobel Prize was awarded for this work to Cornell and Wieman.

Still larger condensate was achieved where 10^{8} hydrogenatoms were condensed to the same quantum state and matter wave like LASER beam was achieved.

1.8.2.3. A COMPARATIVE STUDY OF MAXWELL-BOLTZMAN[Appendix XXXVI] DISTRIBUTION, FERMI-DIRAC DISTRIBUTION[Appendix XXXVII] AND BOSE-EINSTEIN DISTRIBUTION.

Table.1.8. The three statistical distribution function.

	Maxwell-Boltzmann	Bose-Einstein	Fermi-Dirac
Applies to systems of	Identical distinguishable particles	Identical, indistinguishable particles that donot obey Exclusion Principle	Identical, indistinguishable particles that obey Exclusion Principle
category	Classical	Bosons	Fermions
Property of particles	Any spin, particles are far enough hence wave function do not overlap and they remain distinguishable	Spin = 0,1,2….Wave functions are symmetric to interchange of particle labels and are indistinguishable	Spin = 1/2.3/2,5/2, symmetric to interchange and are distinguishable
Example	Molecules of gas	Photons in a cavity,Phonons in a solid,Liquid Helium at 0K	Free electrons in a metal,White dwarf,
Distribution Function	AExp[-E/kT]	1/[Exp{(E + α)/(kT)} -1]	**1/ [Exp{(E-E F)/(kT)} + 1]**
Properties of distribution	No limit to number of particles per state	No limit to number of particles per state;**More**	One particle per state;**Less particles**

particles per state than fMB at low energies and approaches fMB at high energies;

per state than fMB at low energies and approaches fMB at high energies.

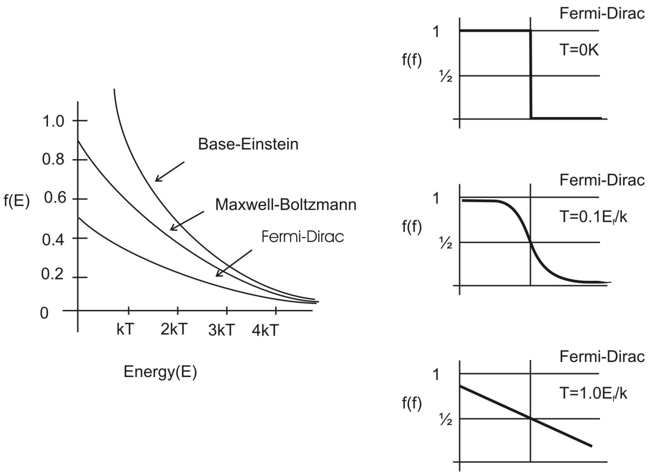


Fig : A comparison of three Distribution Functions.

Fig(1.36) A comparison of three Distribution Functions.

1.8.2.3. COMPARISON BETWEEN BOSE-EINSTEIN AND FERMI-DIRAC DISTRIBUTION.

When indistinguishable particles like photons are brought together they create attraction and there is a tendency of clustering. In an elemental phase space many photons can be accommodated at the same energy state at the same time and there will be no opposition to this tendency of clustering.

But if distinguishable particle like electrons are brought in an elemental phase space then at most two electrons of opposite spin $\pm(1/2)\hbar$ can be accommodated and all other electrons will be repelled. This is called degeneracy pressure.

When a Main Sequence star runs out of its fusion fuel , generation of fusion energy stops and gravitational collapse restarts leading to the formation of White dwarf. The gravitational collapse is finally checked by the degeneracy pressure developed by tightly packed electrons. This is due to the fact that electrons are fermions and not more than two electrons can be accommodated in an elemental phase space.

The phase space has three spatial axes (x, y, z) and three momentum axes (p_x , p_y, p_z).

An elemental phase space $= \Delta V = \Delta x.\ \Delta y.\ \Delta z.\ \Delta p_x\ .\ \Delta p_y\ .\Delta p_z$

But $\Delta x.\ \Delta p_x > \hbar$,

$\Delta y.\ \Delta p_y > \hbar$,

$\Delta z.\ \Delta p_z > \hbar$, from Heisenberg Uncertainty Principle.

Therefore $\Delta V \geq \hbar^3$,

Every elemental phase space has a spatial coordinate x_n , y_n , z_n and momentum coordinates are p_{xn} , p_{yn} .p_{zn} .

In circuit theory we have studied Laplace Transform and in Signal Processing we have studied Fourier Transform. In both cases we have studied the Time Domain Representation and Frequency Domain Representation of the signals.

In an exactly identical manner we have a transformation from Real Space to Phase Space in case of Atomic electrons

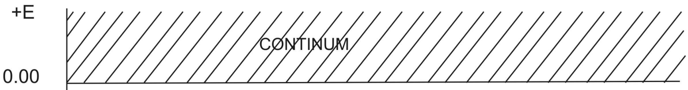
In Real Space, the orbiting electrons have sets of four quantum numbers (n, l, m, s).

In Phase Space, every elemental phase space is filled with opposite spin electrons. With every elemental phase space , we have a set of four quantum numbers associated $(n)_N$, $(l)_N$ $(m)_N$ $(\pm 1/2)_N$. Therefore no two sets of quantum numbers are identical This gives rise to Pauli's Exclusion Principle which states that no two electrons can have all the four quantum numbers identical. In a given phase space, the two opposite spin electrons have identical n , l. m but opposite spin . Thus no two electrons are ever identical in all four quantum numbers.

Sec(1.8.1) indicates that electrons have $2n^2$ -fold degeneracy in nth Shell. As we saw this was incorrect because we had taken non-relativistic Hamiltonian. If relativistic Hamiltonian is considered then we have:

$$\mathbf{E= -13.6eV/(n+l+1)\ 2 \quad 1.81}$$

. Through this equation n-fold degeneracy is removed by taking into account the eccentricity of the ellipsoidal subshell in which the electrons of a given “l” are placed. Thus partial removal of degeneracy is accounted for as shown in Fig(1.37).



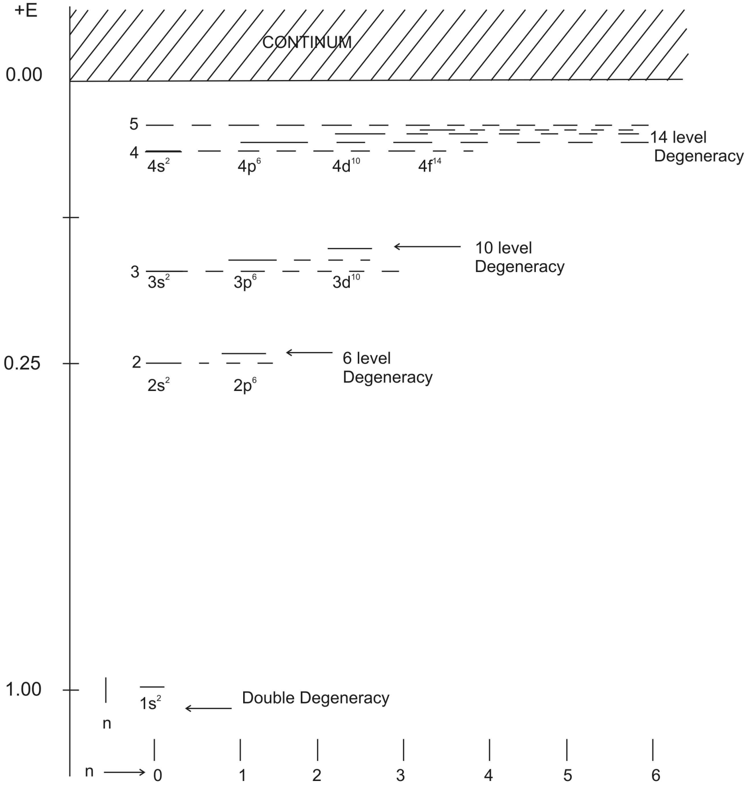


Fig : Partial removal of degeneracy. This removal is due to the orbital shape incorporated in azimuthal quantum number

Figure 1.37. Partial temoval of degeneracy. This removal is due to the orbital shape incorporated in azimuthial quantum number.

When n = 3 then l = 0 (perfectly spherical orbit), 1(ellipsoidal orbit), 2(elongated orbit). These three azimuthial numbers cause the degenerate levels to split apart. Thus three fold degeneracy is removed.

When the atom is put in a magnetic field then remaining 6 fold degeneracy is removed. This is due to the magnetic moment caused due to orbital angular momentum and spin angular momentum. Magnetic moment is vector quantity. Hence it is positive when aligned with the Magnetic field and negative when it is anti-parallel.Let us consider the sub shell $n=3$ and $l=2$. In this subshell each electron has orbital angular momentum of $Lz=\hbar\sqrt{6}$.

Projections of the orbital angular momentum on the magnetic field are $2\hbar$, $1\hbar$,

$0\hbar$, $-1\hbar$, $-2\hbar$. Magnetic interaction will remove five-fold degeneracy.

Total spin angular momentum is $\hbar\sqrt{((1/2)(3/2))}$. This has two projections

There are 10 electrons in this sub shell ($n=3$, $l=2$) : 2(opposite spins) at $m= -2$, 2(opposite spins) at $m= -1$, 2(opposite spins) at $m= 0$, 2(opposite spins) at $m=1$, 2(opposite spins) at $m=2$.. When a magnetic field is applied these 10 electrons interact with the applied magnetic field to split in 10 discrete levels.

The z-axis projection of L and S give a total projection of $Jz=Lz + Sz$.

Hence the 10 electrons produce a projection of:

$(2\pm1/2)\hbar$, $(1\pm1/2)\hbar$, $(0\pm1/2)\hbar$, $(-1\pm1/2)\hbar$, $(-2\pm1/2)\hbar$ on the z-axis. Thus we have 10 discrete and separate magnetic moments resulting into 10 discrete and separate energy levels as a result of the interaction among the applied magnetic field and the magnetic moments. Thus total degeneracy, as shown in Fig(1.35), is removed with the application of the magnetic field. It is because of the removal of degeneracy that we see many more closely spaced linesin the line spectra of an elemental gas than that expected from Bohr’s model.

Stern-Gerlach experiment shown in Figure(1.38) demonstrates the effect of an applied magnetic field upon non-polarized atomic beam.

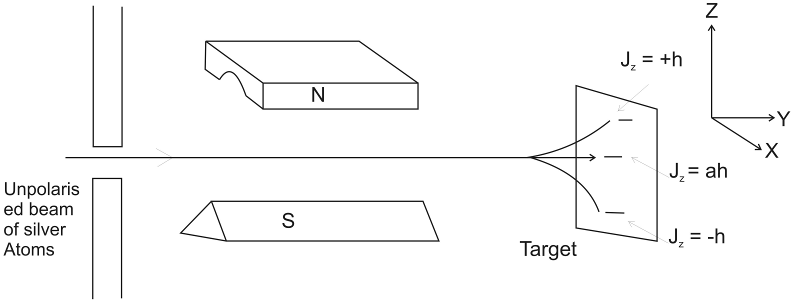


Fig :- Stern - Gerlach Experiment

Figure(1.38) Stern-Gerlach Experiment.

As shown in the Fig (1.38) an unpolarized atom beam passes through a highly non-uniform magnetic field.

Magnetic field is directed in Z direction. The atoms have a net angular momentum of $\hbar\sqrt{2} = J$.

These angular momenta have three possible projections on Z axis:

$J_{z1} = +\hbar$

$J_{z0} = 0\hbar$

$J_{z-1} = -\hbar$

The three projections have three magnetic moments.

Magnetic Moments are: $\mu_z = gJ_z = +g\,\hbar = \mu_{+1}$

$0 = \mu_0$

$-g\,\hbar = \mu_{-1}$

The non-uniform magnetic field creates a magnetic field gradient which interacts with three different kinds of magnetic moments to create 3 different beams of atoms. The scattering forces, which split the beam, are:

$F_{+1} = (\mu_{+1})[\partial H/\partial z]$

$F_0 = 0$

$F_{-1} = (\mu_{-1})[\partial H/\partial z]$

F_{+1} causes a portion of beam to be bent in positive Z direction.

F_{-1} causes another portion of beam to be bent in negative Z direction.

F_0 causes no beam bending.

Thus magnetic field interaction causes the un-polarized beam to split in three distinct portions hitting the target at three separate zones. This establishes that magnetic quantum number is associated with electrons.

1.8.3. BRIEF REVIEW OF THE QUANTUM NATURE OF BOUNDED ELECTRONS.

In this chapter we have established that

1. the bounded electrons have four quantum numbers namely n, l, m and s;
2. for any two electrons all the four quantum numbers can never be identical;

The first proposition establishes the electrons as distinguishable particles. Because they are distinguishable, they follow Fermi-Dirac Statistics. This results in anti-symmetric wave function for electrons which compels electrons to follow Pauli – Exclusion Principle.

SSPD_CHAPTER 1_Part10_ ELECTRON IN A SINGLE CRYSTAL SOLID AND BAND THEORY OF SOLID. SSPD_Chapter 1_Part 10_This section introduces the Band Theory of Solids. Electrons in free space occupy a continuum of states whereas in solid crystalline lattice electrons are permitted to occupy only certain bands of energy states alternated by forbidden zones of energy also referred to as Band Gap.

SSPD_CHAPTER 1_Part10_ ELECTRON IN A SINGLE CRYSTAL SOLID AND BAND THEORY OF SOLID.

All solids are present as crystals: either as single crystal or polycrystal or as amorphous.

In gases the molecules are independent of one another and they have no definite shape and no definite volume. They take the shape of the vessel and occupy the total volume of the vessel..

Liquids have definite volume but no definite shape. They take the shape of the vessel they occupy.

Solids are of definite shape and definite volume. In single crystal solids an unit cell of atoms are repeated in all three dimensions to generate the single crystal lattice structure. The unit cells can be cube, face-centered-cube(fcc), body-centered –cube (bcc), rhombic structure or tetrahedral structure.

In single crystals there is long range periodicity in one given orientation. In poly-crystals the periodicity gets interrupted at a short distance of micrometer. Thus poly-crystal is divided in domains. Each domain has its own orientation of periodicity.

Amorphous state has still shorter range of periodicity. Amorphous structure effects the energy band diagram also. In amorphous solids, the band edges are not sharply defined.

In Figure 1.39.a we illustrate the Single Crystal Structure. Silicon Single Crystal is tetrahedaral structure with 4 electrons in valence orbit completing co-valent bonds with 4 electrons of 4 Si atoms at the corner of the central Si atom. But at the surface of the crystal Si, atoms are not able to complete the co-valent bonds. The incomplete bonds, marked by broken lines, are referred to as 'Dangling Bonds'. The dangling bonds give rise to surface states . These surface states play a crucial role in Solid State Devices. They can give rise to undesirable leakage current. They can give rise to fluctuating channel conductivity in MOSFET devices. The fluctuating channel conductivity gives rise to Flicker Noise or 1/f noise. These dangling bonds have to be passivated by growing SiO_2 layer. Hence Silicon Dioxide is referred to as Passivation Layer.

In Figure 1.39.b., we illustrate the poly crystal structure. Here we have short range single crystal domains separated by domain boundaries. Each domain has its own crystal orientation. Domain boundary gives rise to scattering and it lowers the mobility of the electron hole carriers. Polycrystal Si is not suitable for producing stable and reproducible devices. But in recent days Poly-Si have been used in selective manner in BJT and in MOSFET to improve the performance of the device as we have seen in the Prologue (The Journey of IC Technology).

In Figure 1.39.c. we illustrate the amorphous Silicon. As seen they have very short range periodicity and they have many dangling bonds which will give rise to serious problems of reproducibility and stability in device fabrication.

In Figure 1.39.d. hydrogenated amorphous Silicon is shown. Here the dangling bonds have been terminated by Hydrogen co-valent bonds . Saturating the dangling bonds with hydrogenation improves the prospect of using H:a-Si in device fabrication.

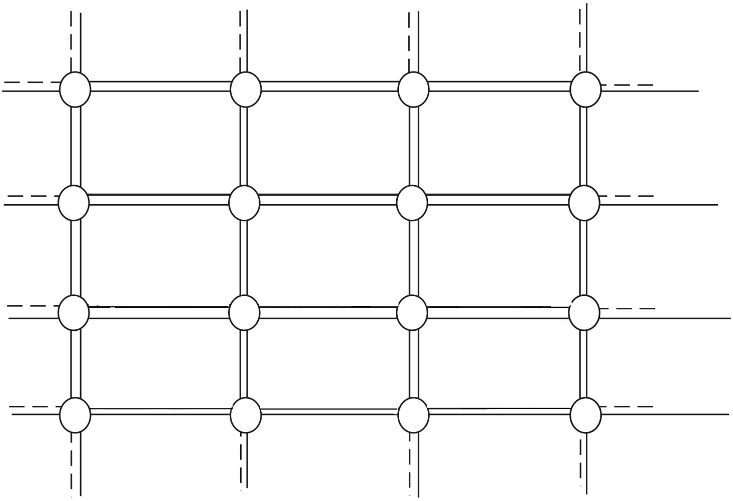


Figure 1.39.a.- Single Crystal

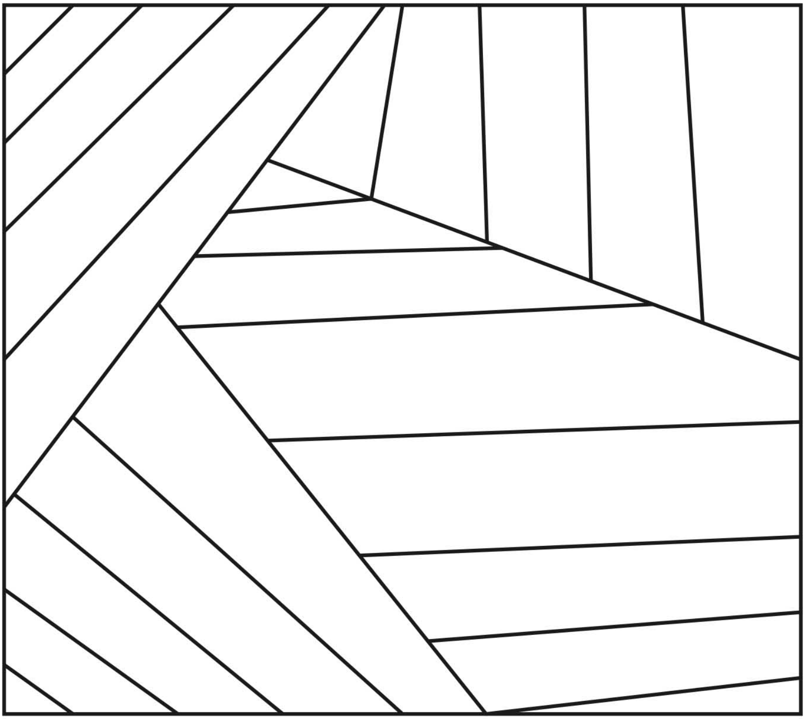


FIGURE 1.39.b. :- Polycrystal

In Figure 1.39.c. we see the unsaturated dangling bonds of Si Atoms. This makes the use of a:Si very unreliable in device fabrication.

In Figure 1.39.d. the same unsaturated dangling bonds have been terminated by Hydrogen valence-bond. This enhances the use of H:a-Si in device fabrication.

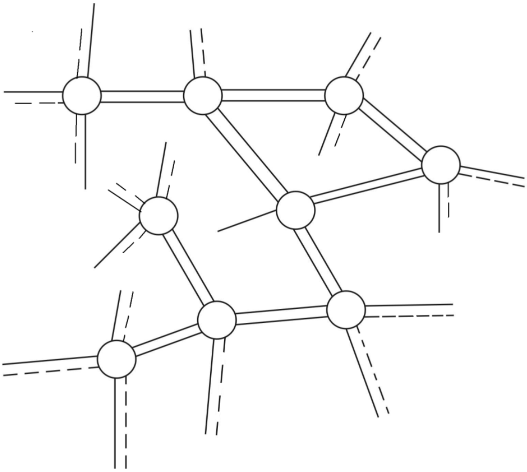
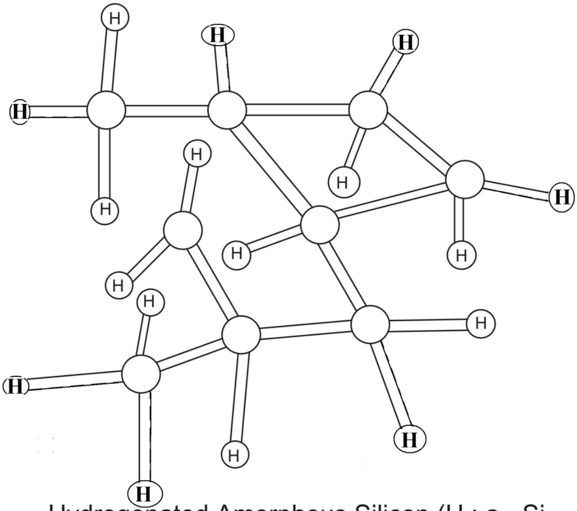


Fig.1.39.c.Amorphous Si(a:Si)



Hydrogenated Amorphous Silicon (H : a - Si

Figure.1.39.d

Figure(1.39) The Three forms of crystalline

structure: Single Crystal, Poly-crystal and Amorphous and the hydrogenated a-Si.

In a later Chapter, crystalline structure will be discussed in detail. Here we are concerned with Energy Band Diagram of single crystal.

An electron in free space can occupy a continuum of energy states given by the equation:

$$\mathbf{E = V(r) + p\ 2\ /(2m)\quad 1.81}$$

But an electron, orbiting around the nucleus of Hydrogen Atom, is permitted to have only discrete energy states as shown in Fig.(1.40 a).

Equation 1.81 states that when electron is in free space it occupies a continuum of energy states as shown by the parabolic solid line in Figure 1.40.a.

But when the same electron is confined to a potential well, as it is in an Hydrogen atom, then it occupies only discrete energy states expressed by the equation:

$$E_n = \frac{-13.6eV}{n^2}$$

This equation was discussed in a previous section Part 7 as Eq 1.35. In Figure 1.40.a it is marked as -13.6eV, -3.4eV, -1.51eV, -0.85eV and -0.544eV for principal quantum numbers n = 1, 2, 3, 4, and 5.

When atoms are arranged into a solid structure, valence and conduction electrons can occupy closely spaced permissible energy states. These permissible energy states are so closely spaced that they appear as a continuous band of permissible energy states. These energy bands are interrupted with energy band gaps. The energy band gaps are forbidden energy states where electrons are not permitted.

The discrete energy states of an electron in a single atom and the permissible energy bands of valence and conduction electrons in a single crystal solid – conductor and insulator are shown in Figure (1.40a,b and c).

Figure 1.40.b. gives the energy band diagram of a conducting solid. As seen in the Figure, conduction band is partially filled and the upper edge of the partially filled band is Fermi Energy Level (E_F). This means Fermi Energy Level is located within the the conduction band. Lower edge of the conduction band is called E_C and upper edge of the valence band is E_V. Because of available empty energy states in the conduction band, a conductor supports an electric drift current. This is why it is called a conductor.

Figure 1.40.c gives the energy band diagram of an insulator . As seen from the diagram, conduction band is fully empty and valence band is completely filled up. Fermi Energy Level lies exactly in the middle of the Band Gap. If the band gap is greater than 5eV it is a perfect insulator but if $E_F < 5$eV then it is a semiconductor above 77Kelvin and below 77Kelvin it continues to be insulator.

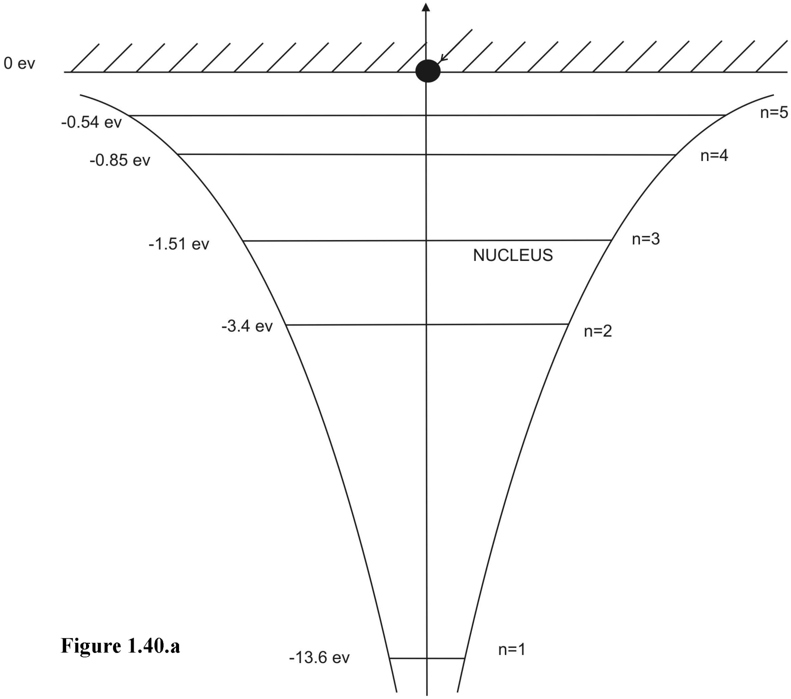


Figure 1.40.a

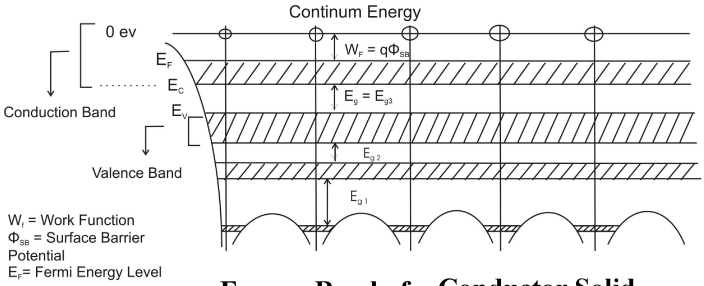
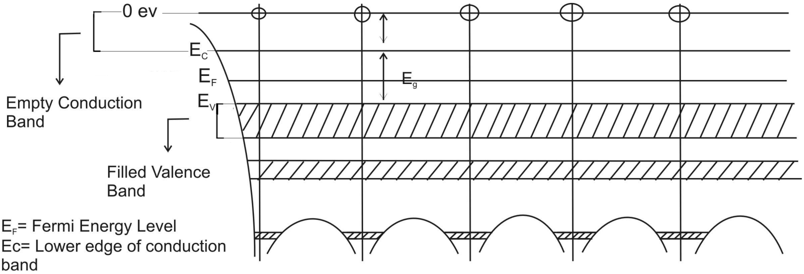


Figure 1.40.b **Energy Band of a Conductor Solid.**



E_F= Fermi Energy Level
Ec= Lower edge of conduction band
Ev= Upper edge of valence band
0 < Eg < 5ev--------Semiconductors
Eg > 5ev----------Insulator

Energy Band Diagram of a **insulator** solid.
Figure 1.40.c

SSPD_CHAPTER 1_Part10_ Continued_The theoretical basis of Band Theory of Solids

Chapter1_Part10_continuation gives the theoretical explanation for the Band Theory of Solids. This part invokes Heisenberg Uncertainty Principle and Pauli's Exclusion Principle to explain the same.

SSPD_Chapter 1_Part 10_continued_EXPLANATION OF ENERGY BAND THEORY BASED ON HEISENBERG’S UNCERTAINTY PRINCIPLE.

Heisenberg Uncertainty Principle states:

$$\Delta x.\Delta p_x \geq \hbar/2$$

$$\Delta t.\Delta E \geq \hbar/2 \tag{1.82}$$

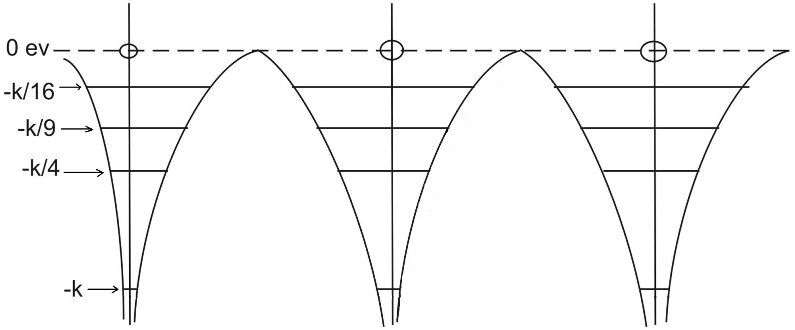
In a single Hydrogen Atom, when an electron jumps from the ground state to excited state then electron stays in the excited state for 10 nanoseconds. On atomic time scale this life-time in excited state is considered to be long. By Hiesenberg’s Uncertainty Principle a large temporal uncertainty in excited state makes the energy level less uncertain. The excited states have uncertainty of only 10^{-7}eV. This means the energy spread has a magnitude of 10^{-7}eV. Hence in single atoms all the states are discrete.

In a crystalline lattice on the contrary, the electrons stay at the excited state only for femtoseconds (10^{-15}

seconds). Because of this very reduced temporal uncertainty in the lifetime of excited state a very large uncertainty, almost of the order of 1eV, creeps in the excited energy state. This leads to spreading of discrete energy states in a single atom to energy bands of the order of 1eV in a single crystal solid.

Why is the uncertainty in the life-time at higher energy states in a solid so minimal?

In a single atom, an electron is trapped in an infinite potential well hence we have well defined discrete energy states as permissible energy states. On the contrary in a crystalline solid, valence electrons are trapped in an array of finite potential wells and the potential walls dividing the potential wells are of finite thickness. This results in a finite quantum-mechanical tunneling probability of excited state electrons . This tunneling probability is higher for higher excited states. Because of this quantum mechanical tunneling phenomena, electrons are highly unstable at the excited states and they immediately return to the ground state. This explains the negligible uncertainty or negligible spread in the lifetime of excited electrons. Smaller is the spread in the lifetime larger is the energy spread of the excited state. As we go to outer orbits, meaning by as we go to higher energy states we have broader energy bands as seen in Fig (1.41.b & c). The innermost energy bands are the narrowest and the outermost are the broadest.



Isolated Atom, its corresponding potential well and the discrete energy level of the electron trapped in such a potential well.(Atoms are well separated and no mutual interaction)

Figure 1.41.a.

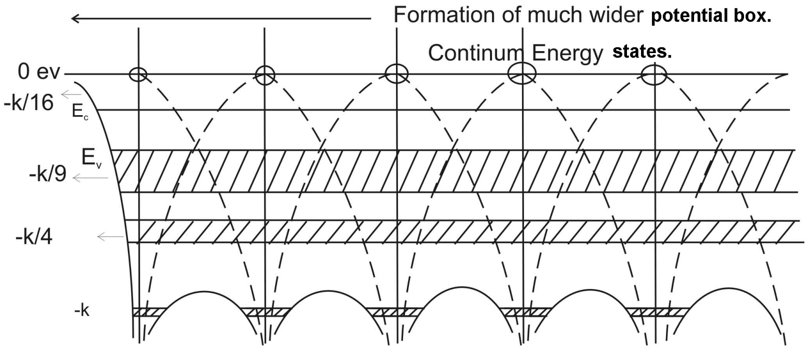


Figure1.41.b.

In a single crystal structure, the array of potential wells superimpose to form the much wider potential well bounded by the surface potential

Figure 1.41. Comparative study of the potential well of an isolated atom and the much wider potential box created by a crystalline solid.

If there is 1-D array of atoms well separated as shown in Figure 1.41.a, the electrons continue to occupy discrete energy states given by the formula :

$$E_n = \frac{-K}{n^2}$$

But as the inter-atomic distances are reduced and electrons in individual potential wells start interacting the discrete energy states start spreading out in closely spaced energy levels. This cluster of closely spaced energy states appear as a permissible band of energy states.

In a solid if electrons are excited to an energy level greater than 0eV then the electron escapes out of the solid and is now present in free space where it can occupy a continuum of energy states. Hence we show a continuum of energy state above 0 eV.

SSPD_Chapter 1_Part10_continued_EXPLANATION OF ENERGY BAND THEORY ON THE BASIS OF PAULI'S EXCLUSION PRINCIPLE.

A single Atom and a 3-D array of atoms arranged in a crystalline solid have a fundamental difference.

A single atom has a single potential well which acts as the bounded space surrounding the nucleus. Within this bounded space only those electrons are permitted which fulfill the boundary conditions for standing waves. Since these electrons are fermions they obey Pauli's Exclusion Principle. In an elemental phase space only two electrons of opposite spin can be accommodated. What this means is that no two electrons can be at the same energy level at the same spatial coordinate. If two electrons are at the same energy level, spatially they must spread out. If spatially they are squeezed to the same coordinates then they must spread out in energy sense.

When there are two atoms well separated, electrons in their respective host atoms can continue to be in the ground state and the two ground states will be equal. Here the two systems are non-interacting and their phase spaces are well separated.

In a crystalline solid, the 3-D array of potential wells constitute a common phase space bounded by the surface potential barriers of the solid and having a periodic potential variation throughout the three dimensional phase space. Now the electrons of the respective atoms belong to the common phase space and these electrons need to fulfill the boundary condition of standing waves as set by the common phase space and not by the potential well of individual atoms. The individual potential wells are drastically modified when the atoms are three dimensionally arranged into crystalline solid as shown in Fig(1.41. b).

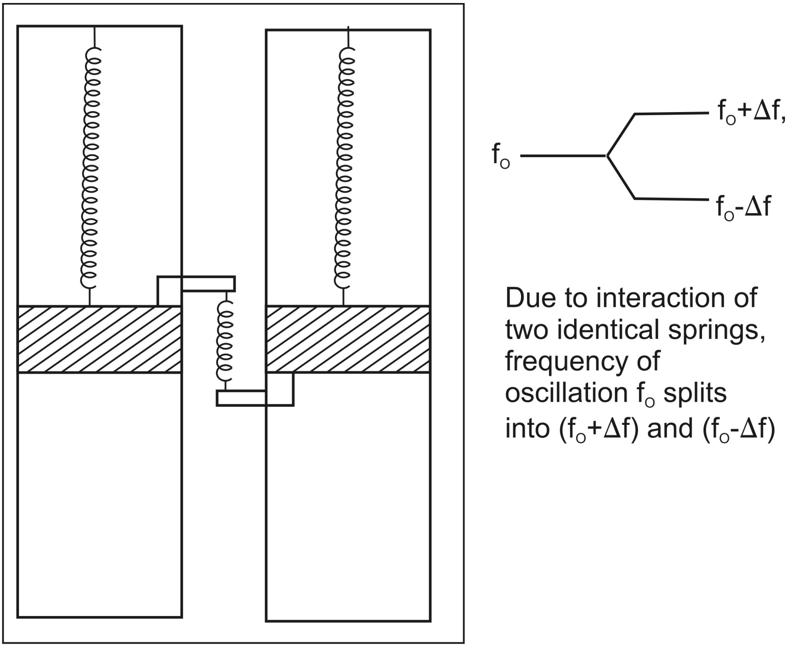
The electrons do not belong to the individual host atoms but belong to the whole lattice space. Electrons instead of being localized are non-localized but their localization is destroyed to definite degree depending upon the orbital position of the electrons.

In a crystalline solid with a large atomic number, the inner orbital electrons are tightly bound to their respective nuclei and hence their localized nature is retained. The energy bands corresponding these electrons are narrow.

As we move outward, the localized nature is progressively destroyed and progressively they are mutually interacting. These leads to wider and wider broadening of the energy bands.

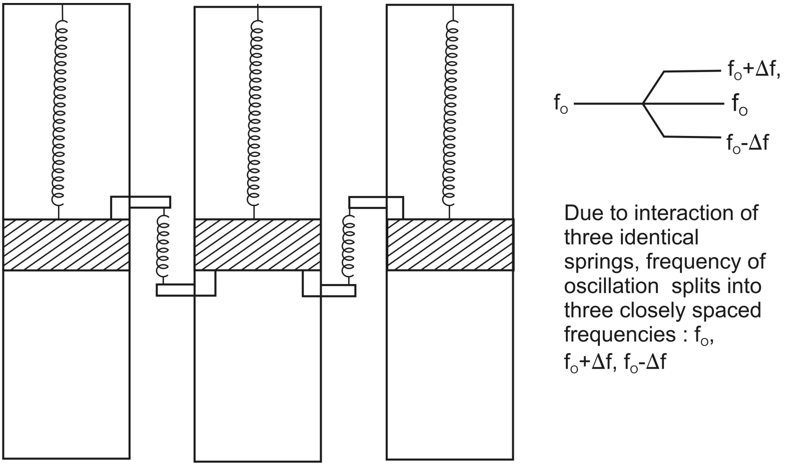
In Figure(1.41), we show the nature of an individual potential well and also the nature of the much wider potential well formed by the combination of the linear array of potential wells. We have also shown the periodic potential variation in the wide potential box caused by the total solid.

Within the single crystal solid, the creation of the much wider potential box and the periodic potential variation within the box are the underlying cause of the energy bands.



Due to interaction of two identical springs, frequency of oscillation f_o splits into $(f_o+\Delta f)$ and $(f_o-\Delta f)$

Figure 1.42.a. The arrangement of two coupled springs.



Due to interaction of three identical springs, frequency of oscillation splits into three closely spaced frequencies : f_o, $f_o+\Delta f$, $f_o-\Delta f$

Figure 1.42. b three coupled springs interaction

Figure 1.42. The effect of two coupled springs and three coupled springs in splitting the individual energy levels and their natural frequencies of vibrations.

In Figure (1.42) we show two identical coupled spring loaded mass system and three identical coupled spring loaded mass system. The coupling is achieved through a massless and weak spring. Each individual spring loaded mass is constrained by frictionless guides. When the main spring loaded masses are uncoupled they exhibit identical frequencies of oscillation ν_0. In two spring loaded mass system coupling results in splitting of the oscillation frequencies into two closely spaced frequencies namely :

$\nu_0 + \Delta\nu$and ν_0-$\Delta\nu$.

In three spring loaded coupled system the central frequency splits into three closely spaced frequencies and in N coupled system the central frequency splits into N closely spaced frequencies. If N is very large then this splitting results into a continuous band of frequencies.

If we consider the frequency of the loaded spring to be analogous to the ground energy state of the electron in a hydrogen atom then two interacting hydrogen atoms will lead to the discrete ground energy state splitting into two energy states and the same splits into three energy states when three atoms are mutually interacting. By corollary a discrete ground energy state splits into N closely spaced energy states when there is a linear array of N atoms close enough to be mutually interacting. This N closely spaced energy states appears as a band of permissible energy states.

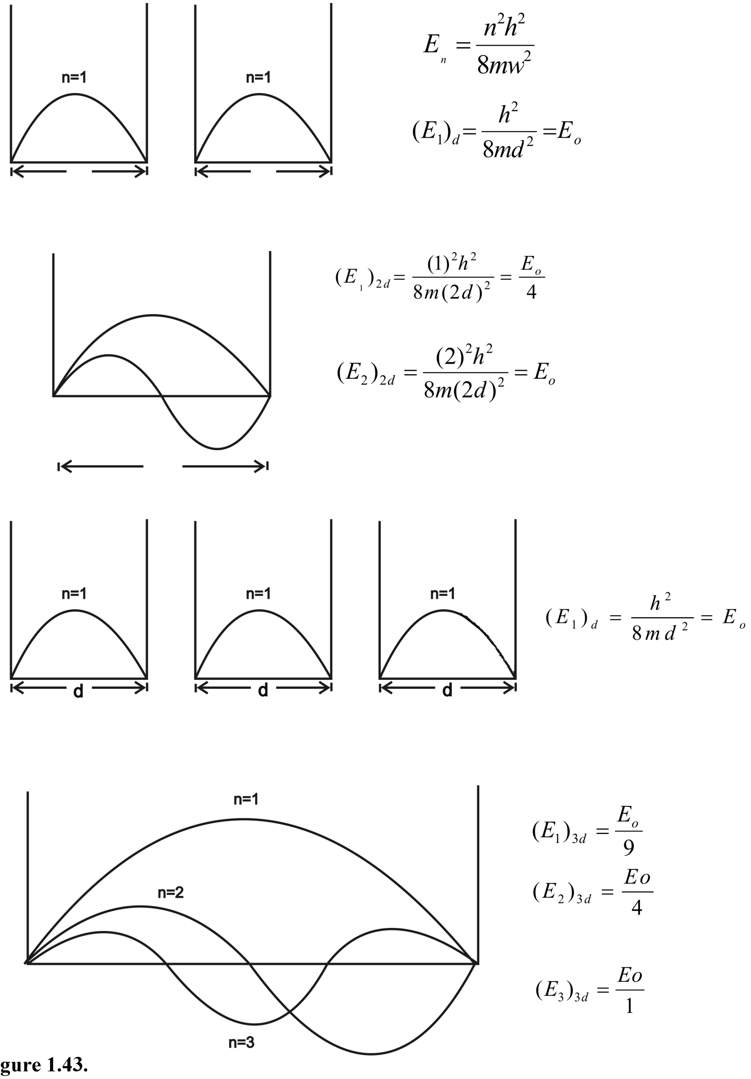

n=1
n=1
$E_n = \frac{n^2h^2}{8mw^2}$
$(E_1)_d = \frac{h^2}{8md^2} = E_o$
$(E_1)_{2d} = \frac{(1)^2h^2}{8m(2d)^2} = \frac{E_o}{4}$
$(E_2)_{2d} = \frac{(2)^2h^2}{8m(2d)^2} = E_o$
n=1
n=1
n=1
d
d
d
$(E_1)_d = \frac{h^2}{8md^2} = E_o$
n=1
n=2
n=3
$(E_1)_{3d} = \frac{E_o}{9}$
$(E_2)_{3d} = \frac{Eo}{4}$
$(E_3)_{3d} = \frac{Eo}{1}$

Figure 1.43.

Fig: Merging of two potential wells into potential well and merging of three potential wells into one potential well.

Figure 1.43. Merging of two potential wells into one potential well and merging of three potential wells into one potential well.

Let us examine two isolated atoms each with a diameter of d Å. The symmetric matter wave of principal quantum number n $=1$ corresponds to $E_0= h_2/(8md_2)$.

The symmetric matter wave of each atom will combine to give a symmetric wave and anti–symmetric wave belonging to a combined potential well of dimension 2d Å. Hence the resultant symmetric wave will have energy $E_1= h_2/(32md_2) = E_0/4$ and the anti-symmetric wave will have energy $E_2 = h_2/(8md_2)= E_0$. Thus two atoms interaction result in the symmetric wave energy E_0 splitting into (anti-symmetric)E_0 and (symmetric)$E_0/4$ energy states.

Let us examine three isolated atoms. The lowest symmetric wave corresponds to E_0 belonging to the principal quantum number $n=1$. When they are brought close enough to be interacting then the symmetric wave will combine to give three possible combinations: the lowest will be a symmetric wave of $n=1$, the anti-symmetric wave of $n=2$ and the next highest symmetric wave of $n=3$, all belonging to a combined potential well of dimension 3d Å. These three matter waves have energies :

$E_1= h_2/(8m(3d)_2)= E_0/9$; $E_2= 4h_2/(8m(3d)_2) =(4/9)E_0$; $E_3 =9 h_2/(8m(3d)_2)= E_0$;

Thus 3 atoms interaction cause the first energy state split into 3 energy states. By the same logic we can conclude that N atoms will interact to split the ground state E_0 into N energy states where lowest energy level would be E_0/N and highest energy state would be E_0 .

This simple model based on ‘electron in a potential well’ explains how a discrete energy state splits into N discrete energy statesbut this model completely fails to show as to how the continuous band results.

This model is too simplistic and does not correspond to the reality. We have assumed a flat bottomed potential well where as in fact it is an undulating surface due to periodic potential variation as shown in Fig(1.41.b). A more accurate model is Kronig Penny Model which is the topic of discussion in the next section

SSPD_CHAPTER 1_Part10_ Concluded_Kronig-Penney Model
Chapter 1_Part10_Conclusion describes the theoretical basis of Energy Band Theory of Solids by analyzing Kronig-Panney Model which idealizes 1-D crystalline array and determines the Schrodinger Equation Solution in 1-D array of square wells.

SSPD_Chapter 1_Part10_concluded_ ENERGY BAND THEORY IN A SOLID BASED ON KRONIG-PENNEY MODEL.

Kronig-Penney model is 1-D array of square wells as shown in Figure 1.44.. This is 1-D idealization of a linear array of atoms in a single crystal lattice structure. The solution of Schrodinger Equation using this array of Square wells becomes more tractable and it still brings out the important features of the quantum behavior of electrons in real life crystalline periodic lattice.

There are four assumptions in Kronig-Penney Model analysis namely:

1. Electron interaction with the core is purely coulombic ;
2. Electron to electron interaction is precluded;
3. Non-ideal effects, such as collisions with the lattice and the presence of impurities, are neglected;
4. Atoms are fixed in position whereas they are

having thermal vibrations.

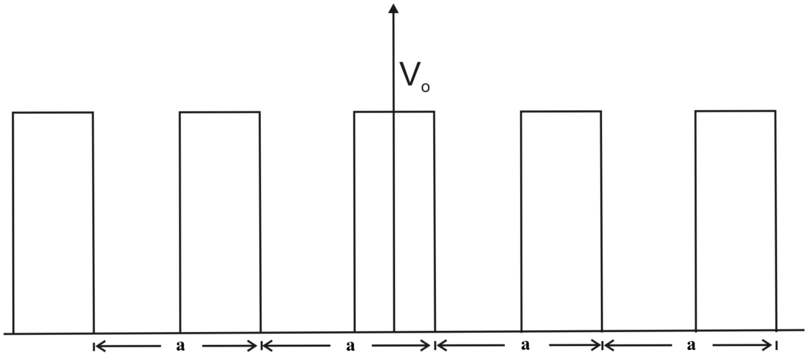


Figure 1.44. Kronig-Penny Model of a linear array of atoms in a single crystal solid.

The solution of Schrodinger Equation can be arrived at mathematically but for simplicity of presentation we will discuss the problem in qualitative terms only.

Study of electron in a crystalline structure is really the study of an electron in a periodically varying potential field. For simplicity of analysis we assume a linear array of

atoms . The crystal length is L cm. Let Z-axis be the longitudinal axis and let the crystal be repeated along the Z-axis with a period of L cm from - ∞ to + ∞ . Along X-axis and Y-axis it is of infinite length. So we have a semi-infinite crystal of finite length L cm in Z-axis. For the ease of calculation we assume that crystal is repeated along z-axis at L cm.

Since we have assumed a periodic crystalline structure along Z axis therefore the solution of the Schrodinger Equation is applicable only in the bulk and not at the boundaries of the crystal.

We will assume that L cm = 1cm $= 1 \times 10^{-2}$m. The crystal structure is referred to as the lattice. The atoms of the crystal are referred to as the lattice centers. The distance between two consecutive lattice centers is referred to as the lattice constant ‘a’ Å. A typical lattice constant is 2 Å. Therefore the linear array contains L/a = 1×10^{-2}m/2×10^{-10}m = 5×10^{7} atoms in one period. Let this number be N i.e. N= 5×10^{7} .

We have a periodically varying potential field along the linear array with a periodicity of ‘a’ Å hence the Fourier Series Expansion of the potential is:

$$V(z) = \sum_{m=-\infty}^{\infty} V_m Exp[jm(\frac{2\pi}{a})z]$$

1.83

The potential field has a period of ‘a’ Å hence 2π/a is the fundamental periodicity and the harmonics are 2(2π/a) , 3(2π/a) , 4(2π/a) ……………………m(2π/a)

In the periodic potential field following is the

Schrodinger Equation for time independent part:

$\partial^2 \psi/\partial z^2 + [\{2m(E-V(z))\}/\hbar^2]\psi = 0$………… 1.84

If we had assumed that our very wide potential well was flat bottomed with V(z) = 0 everywhere along the potential box then the solution of the Schrodinger Equation would be a progressive wave as would be obtained for free space:

That is $\psi(z,t) = \psi_0 \mathrm{Exp}[j(\omega t-kz)]$ ……………… 1.85

An electron in 1-D potential box of width dimension L will follow the boundary conditions as given in the table and as shown in Figure 1.13(electron in a infinite potential well).

	k1	k2	k3	k4
L	λ/2	λ	3λ/2	2λ
λ	2L	L	2L/3	L/2
k = 2π/λ	2π/(2L)	2π/(L)	2π/(2L/3)	2π/(L/2)
k	1 × (π/L)	2 × (π/L)	3 × (π/L)	4 × (π/L)

Therefore $k_n = n \times (\pi/L) =$ wave vector of matter wave;

From Chapter1_part 3 and Chapter_part 6 we know that:

$k_z = \sqrt{(2mE)}/\hbar = k$; 1.86

Since the linear array has a period of L cm hence the matter wave has wave vector which is an integer multiple of (π/L) as shown in the Table above.

From Eq.(1.86) we obtain:

$E = (\hbar k_z)^2/(2m) = p_z^2/(2m)$ 1.87

Equation(1.87) gives us the energy distribution with respect to momentum vector (p_z) or wave vector (k_z) . This distribution is continuous and parabolic and is shown in Figure(1.45). This is what would be expected for electron in free space. We get this result here because a very wide potential well with a flat bottom acts as a free and unbounded space for electron under consideration.

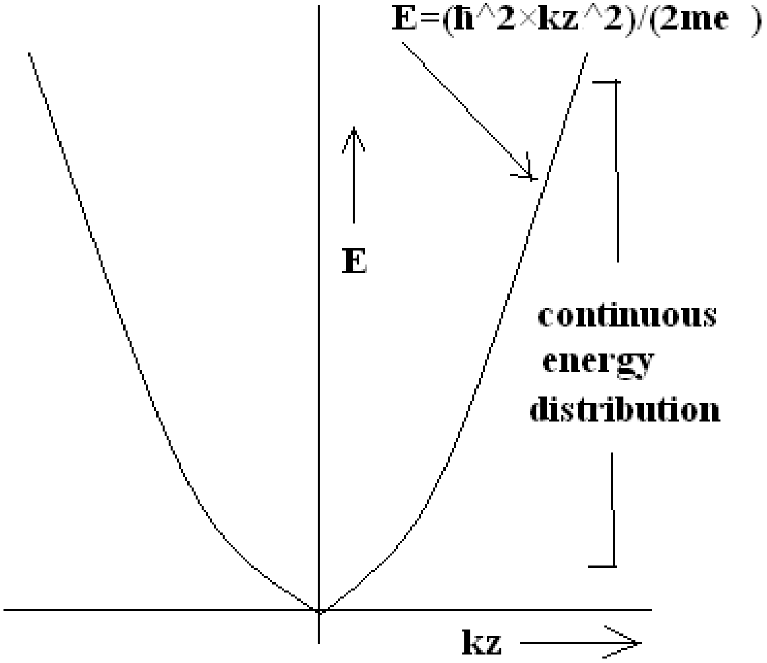


Figure 1.45.Energy distribution of electron in wide potential box.

Figure. 1.45. Energy Distribution for electron in a very wide potential box.

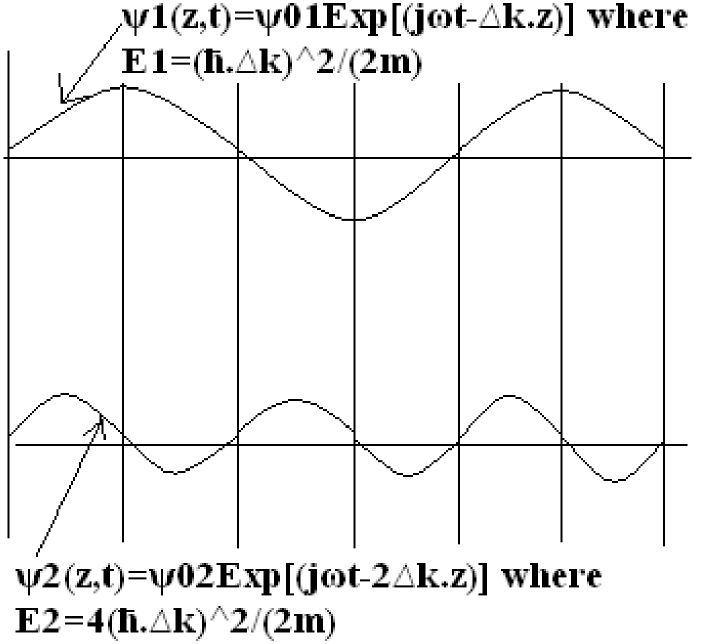

ψ1(z,t)=ψ01Exp[(jωt-Δk.z)] where
E1=(ħ.Δk)^2/(2m)
ψ2(z,t)=ψ02Exp[(jωt-2Δk.z)] where
E2=4(ħ.Δk)^2/(2m)

Figure 1.46.Schrodinger Equation Solution in a wide potential box with a flat bottom.

Figure 1.46. Schrodinger Equation Solution in a wide potential box with a flat bottom.

In Figure 1.46 the progressive wave solutions of the Schrodinger Equation is shown when applied to flat bottomed potential box. The solutions are:

First principal quantum number:
$\psi_1(z,t) = \psi_{01}Exp[(j\omega t\text{-}\Delta k.z)]$ where $E_1 = (\hbar.\Delta k)^2/(2m)$;

Second principal quantum number:
$\psi_2(z,t) = \psi_{02}Exp[(j\omega t\text{-}2\Delta k.z)]$ where $E_2 = 4(\hbar.\Delta k)^2/$

(2m);

……

nth principal quantum number:
$\psi_n(z,t)=\psi_{0n}\mathrm{Exp}[(j\omega t-n\Delta k.z)]$ where $E_n = n^2.(\hbar.\Delta k)^2/(2m)$;

Here $\Delta k = (\pi/L)$ is a very small quantity therefore E_1 , E_2 ,……. E_n constitute a continuous spectrum of energy.

Since we have semi-infinite crystal of length L along z-axis hence each progressive wave has a periodicity of 2L, L, 2L/3, L/2, ……2L/n

Hence $\psi_n(z) = \psi_n(z+2L/n)$

When we replace a flat bottom by undulating bottom due to periodic potential variation due to lattice centres then we have a modulation effect on the progressive waves. As we had already mentioned in Fig(1.44) the periodicity of this potential variation is ‘a’ Å. Mathematical formulation of this modulation effect is difficult therefore we will give a graphical description of this modulation.

Periodic potential field is created by the lattice centers. Each lattice center is a potential well. If the potential well is of width W then electrons exist as standing waves of wavelengths (already shown in

Figure 1.13)

2W, W, (2/3)W, W/2………

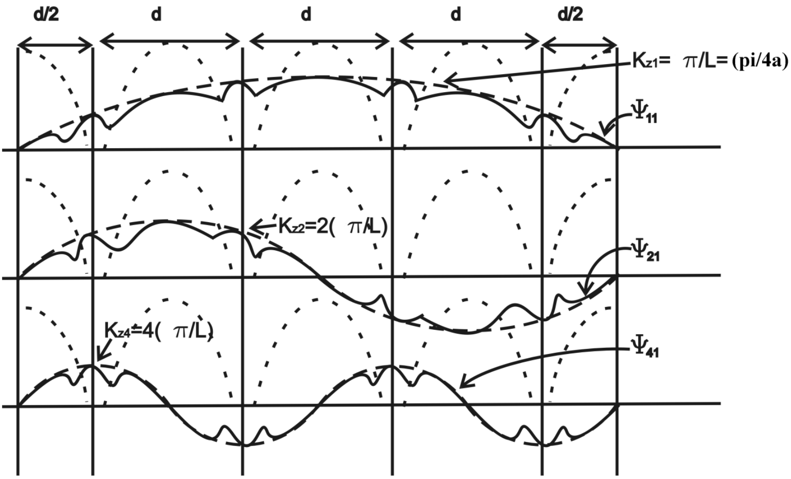


Fig 1.47 Schrodinger Equation Solution in a wide potential box with an undulating bottom for ψ_{11} ,ψ_{21} and ψ_{41} where first subscript gives the wave number and the second subscript gives the level of modulation

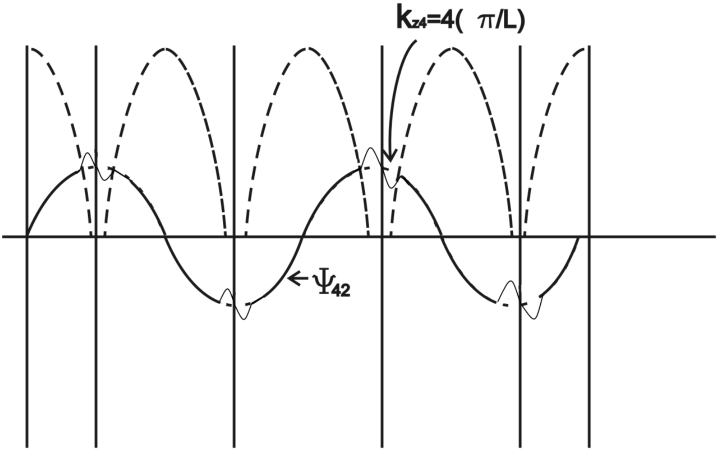


Fig1.48 Schrodinger Equation Solution in a wide potential box with an undulating bottom for Ψ_{42}

Figure 1.48. Schrodinger Equation Solution in a wide potential box with an undulating bottom for ψ 42

In Figure 1.47 and 1.48 we show the modulation effect of the lattice centers. The modulation is caused in the same way as the standing wave is created in the potential wells. Thus in Kronig-Penney Model solution ψ_{nj} there are two quantum numbers:

n is the principle quantum number which decides the wave vector of the electron and which is a multiple of (π/L) hence the wave vector is $k_z = n.(\pi/L)$. This result has been explained in the preceding paragraph.

j is the quantum number which describes the level of modulation.

In Figure(1.47), we describe three matter waves: ψ_{11} , ψ_{21} , ψ_{41} . In Figure 1.48 we show matter wave ψ_{42}

For simplicity of description we have assumed linear array of 4 lattice centers i.e.

$L = 4a$;

We have seen in Part 8_conclusion that at a step voltage there is a reflected wave. When two reflected waves constructively interfere we have a condition of Bragg Reflection or Specular Reflection. *Under this condition forward and reflected wave cancel each other and electron does not exist as a physical entity. Hence the energy states corresponding to this Brag Reflection is forbidden.*

Constructive Interference occurs when:

Path difference of two normally incident waves = $2a = n\lambda$ as we saw in x-ray crystallography.

Destructive Interference occurs when:

Path difference of two normally incident waves = $2a = (n+1/2)\lambda$;

Therefore when the wavelength of matter wave satisfies the condtion:

$\lambda = 2a/n$................. 1.88.

That particular wavelength electron suffers Bragg Reflection.

Bragg reflection was discussed in context of matter wave and its experimental verification of matter wave in Chapter 1_Part 3. Bragg had done X-Ray diffraction using crystalline grating and Davisson-Germer had done electron diffraction again using crystalline grating.

The wave vector of matter wave is : $k_z = 2\pi/\lambda$;

Substituting Eq.(1.88) in wave vector we get:

$k_z = (\pi/a)n$.................. 1.89

Equation 1.89 implies that electrons in a crystalline lattice will not exist at k_z equal to :

$\pm(\pi/a)1$, $\pm(\pi/a)2$, $\pm(\pi/a)3$, $\pm(\pi/a)4$ ....

Bragg brothers had received the Nobel Prize for demonstrating the mirror like reflection or specular reflection of X-Rays from crystal surfaces as described in Chapter 1_Part 3.

From Equation (1.88) and Equation (1.89) we see that matter wave does not exist due to Bragg

Reflection and discontinuity is created at those points:

The points of discontinuity:

Points of discontinuity in E-k diagram at k_n	$k_1 = \pi/a$	$k_2 = 2\pi/a$	$k_3 = 3\pi/a$	$k_4 = 4\pi/a$
But $k_n = 2\pi/\lambda_n$ therefore λ_n of the matter wave	$\lambda_1 = 2a$	$\lambda_2 = a$	$\lambda_3 = 2a/3$	$\lambda_4 = a/2$
In terms of principle quantum number nWhere $L = 4a$	$n\pi/L = \pi/a$	$n\pi/L = 2\pi/a$	$n\pi/L = 3\pi/a$	$n\pi/L = 4\pi/a$
Principle quantum number	4	8	12	16
Corresponding	$\psi_{4,1}$	$\psi_{8,1}$	$\psi_{12,1}$	$\psi_{16,1}$

This implies that these matter waves will not exist. Therefore matter waves at wave vectors : **$\pm(\pi/a)$, $\pm(2\pi/a)$, $\pm(3\pi/a)$, $\pm(4\pi/a)$, ...**

do not exist because they suffer specular reflection or Bragg Reflection.

At these wave vectors of discontinuity, different levels of modulation contribute to bands of energies at wave vectors **$\pm(\pi/a)$, $\pm(2\pi/a)$, $\pm(3\pi/a)$, $\pm(4\pi/a)$, ...**

Hence the non-existence of the matter waves at these points of Bragg Reflection lead to band gaps in otherwise continuous energy distribution with respect to wave vector. Hence continuous E- k diagram shown in Fig(1.45) becomes Energy Band diagram alternated by Energy Gaps as shown in Fig(1.49). The Energy Band Gaps are also referred to as Forbidden Band Gaps.

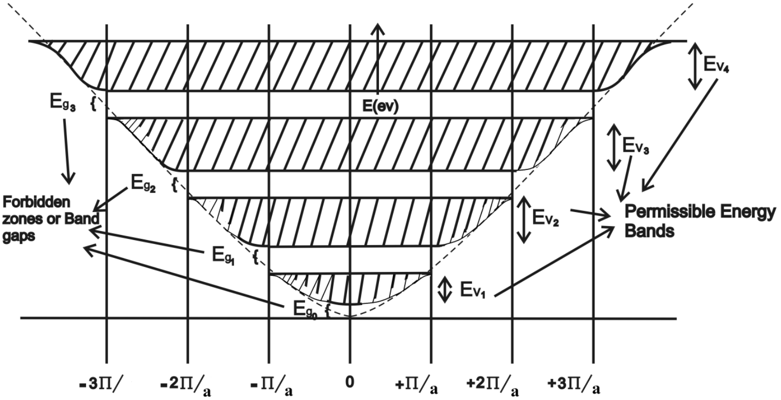


Fig 1.49 Energy Band Diagram of electrons in a crystal solid -the interruption of E-k diagram due to Bragg Reflection

Figure(1.49) Energy Band Diagram of electrons in a single crystal solid.

Through the above discussion we have established qualitatively that electrons in a single crystal solid occupy permissible bands of energy alternated by forbidden band gaps.

Solid State Band Theory has introduced us to a new quantum mechanical phenomena called Bragg reflection.

We have seen that at k= $\pm(\pi/a)$, $\pm(2\pi/a)$, $\pm(3\pi/a)$, $\pm(4\pi/a)$,........

electrons suffer specular reflection which does not permit the independent existence of electrons at those wave vectors or at those momenta vectors since $p = k\hbar$. If we closely examine E-k diagram at the discontinuities we see that E-k curve does not follow the parabolic relationship as it does elsewhere. Hence the behavior of electron at the discontinuities is quite different from its behavior elsewhere. To examine the behavior at the discontinuities we will have to re-examine the Schrodinger Equation in spatial coordinates:

$\partial^2 \psi/\partial z^2 + [\{2m(E-V(z))\}/\hbar^2]\psi = 0$................ 1.90

V(z) is the periodic potential variation;

In an infinite crystal with a period of $L = Na$ where a is the lattice parameter the solution is a progressive wave:

Namely $\psi(z,t) = \psi_0(z)\mathrm{Exp}[j(\omega t - kz)]$

Where $k = \sqrt{[2m\{E-V(z)\}]}/\hbar$

Therefore $E = \hbar^2k^2/(2m_e) + V(z)$;

If we assume an effective mass m_e^* which incorporates the periodic potential variation in itself then we get an expression:

$E = \hbar^2 k^2/(2m_e^*)$............ 1.91

Double differential of E with respect to k gives us the expression of the effective mass:

m e * = ħ ˆ2 /(∂ 2 E/∂ 2 k)…………… 1.92

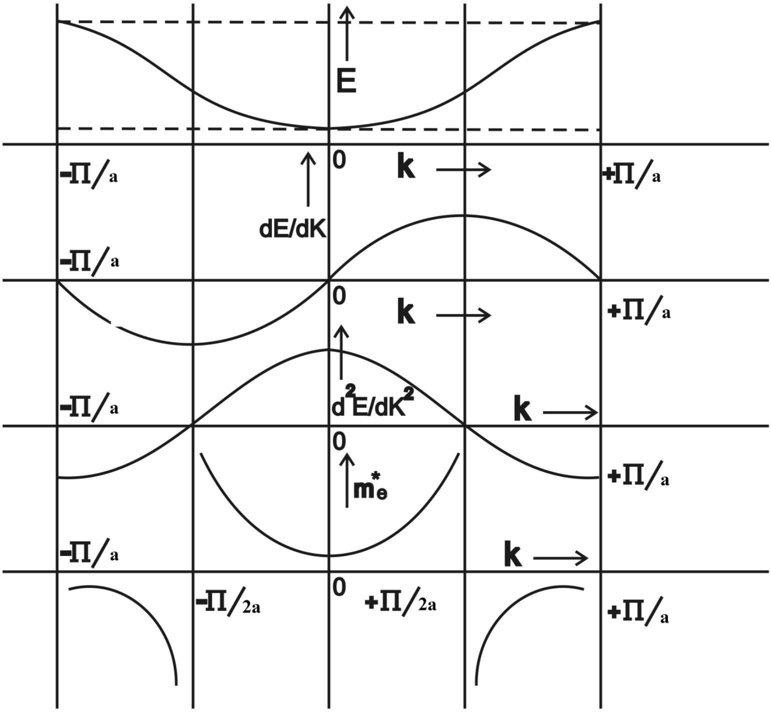


Fig 1.50 Physical interpretation of E-k diagram - the first derivative and the second derivative

Figure 1.50. Physical interpretation of E-k diagram.

The following conclusions result from the inspection of the E-k diagram and its derivatives:

1. Electrons have least effective mass at the bottom of the conduction band;
2. Electrons have negative mass at the top edge of the valence band;
3. Electrons have very heavy effective mass at $\pm(\pi/2a)$ i.e. intermediate values of k. Hence effective mass of electrons in conductor will depend upon the position and location of the FERMI-LEVEL. If E_F is in neighborhood of $\pm(\pi/2a)$ then we obtain heavy Fermionic Metal.

Therefore effective mass is fully dependent on E-k relationship. In Figure(1.50) the first derivative and the second derivative are given. As seen from the diagram, near the lower band edge the second derivative $\partial 2E/\partial 2k$ is higher than the normal value hence at the lower band edge effective mass is lighter than the free space mass. At the upper edge **$\partial 2E/\partial 2k$** is negative hence mass is negative . This gives rise to the phenomenological particle called holes. Holes are the movement of vacancies.

Band theory illustrates a new property of electron namely the property of scattering which is responsible for the resistivity of a conducting solid. Next chapter is devoted to scattering of electrons and resistivity.

Table 1.9 The effective mass m e * of conducting electrons in different metals.

Metal	Ag	Au	Bi	Cu	K	Li	Na	Ni	Pt	Zn
m_e^*/m_e	0.99	1.1	0.047	1.01	1.12	1.28	1.2	28	13	0.85

SSPD_Chapter 1_Part 11_Solid State of Matter
SSPD_Chapter1_ Part11 gives a elaborate classification of Solid State of matter. This first part of Part 11 attempts to explain that electron's debroglie wavelength will decide how it interacts with solids. It is the difference in this de-broglie wavelength which makes conducting electron behave so very different in metals and semi-conductors.This part also describes how particle accelerators are gigantic microscopes which probe innermost recesses of matter.

SSPD_Chapter 1_Part 11_Solid State of Matter

1.10.DEFINITION OF INSULATORS, SEMICONDUCTORS AND INSULATORS.

1.10.1. SIX STATES OF MATTER.

There are six states of matter: gas, liquid, solid, plasma, Bose-Einstein Condensate and Fermi-ionic Condensate.

Gas has no fixed shape or volume. They have the shape and volume according to the vessel they occupy. Inter molecular distances are large and molecules are independent of one another. This state of matter obeys the Ideal Gas Law.

Liquid has a fixed volume but no fixed shape. It takes the shape of the vessel it occupies.

Intermolecular distances are fixed and molecules experience cohesive force with respect to one another.

Solids have a fixed volume and fixed shape. The molecules are arranged in an orderly fashion giving rise to a crystalline structure. Because of the variation in the range of orderliness , the crystalline structure is classified as Single Crystal, Poly-crystal and Amorphous. The range of orderliness is shown in Figure(1.39).

For 300,000 years after the Big-Bang, temperature of the Universe was above 4000K and all matter was in PLASMA STATE. This is a soup of electrons, protons and neutrons. As long as matter is in plasma state, RADIATION dominates. Gravitational attraction is dominated by electromagnetic forces and thus the gravitational accretion is prevented.. As soon as temperature falls below 4000K, plasma recombines to form neutral mass of Hydrogen (70%), Helium(30%) and traces of Lithium. At this point radiation decouples and matter dominates. The decoupled radiation carries the imprint or profile of the matter distribution through out the Universe at the time of de-coupling. This decoupled radiation persists till today in almost its pristine state in which it decoupled from the matter. This decoupled radiation is known as Cosmic Microwave Background Radiation (CMBR). The latest study of CMB by WMAP show that indeed there are hot spots and cold spots in CMB implying that in the remote past matter distribution did contain the unevenness which would eventually become the seeds for the formation of clusters, galaxies, solar systems and planets.

Bose-Einstein Condensate and Fermi-ionic condensate are described in Section (1.15) . They are closely related but Fermi-ionic Condensate formed at a lower temperature than the temperature at which Bose-Einstein Condensate is formed. It manifests a complex of spectacular behavior:

1. It flows through tiny capillaries without experiencing any friction;
2. It climbs in the form of film over the edge of vessels containing it. This phenomena is referred to as ‘film creep’;
3. It spouts in a spectacular way when heated under certain conditions;
4. If it is contained in a rotating container, the content in He-II Phase never rotates along with the container.

1.10.2. SOLID STATE OF MATTER.

With the development of Quantum Mechanics, Band Theory of solid was proposed by Felix Bloch[Appendix XXXXIII]. As already seen in the last section, electrons in a single crystal solid occupy energy bands separated by forbidden zones known as Band Gaps as shown in Figure(1.40) , Figure(1.41) and in Figure(1.49).

The outer band is known as the conduction band and the band just below is valence band. In insulators and semiconductors, the energy band gap between valence and conduction is wide of the order of eV and conduction band is completely empty at low temperatures for both insulators and semiconductors. That is at 77K (liquid Nitrogen temperature) and below, semi-conductor is an insulator. Only near room temperature and by introduction of controlled amount of impurities that semiconductor acquires a certain degree of conductivity. As we will see in semiconductor chapter the effective density of states at the lower edge of conduction band (Nc)

and at the upper edge of the valence band (Nv) is nearly the same as shown in Table(1.10). This is of the order of 10ˆ19 permissiblestates per centimeter cube. Hence as soon as *doping approaches this order of magnitude, number of conducting electrons are comparable to the available energy states which is the criteria for degenerate systems. Hence at that order of magnitude of doping, semiconductor becomes degenerate and behaves like a semi-metal.*

So an INSULATOR can be defined as the solid which has an empty conduction band and a large band gap , of the order of 4eV or more. There electron- hole pair cannot be thermally generated . Hence it remains non-conducting at all temperature.

On the contrary, SEMICONDUCTORS are non conducting and hence insulator below liquid Nitrogen temperature and above liquid Nitrogen temperature they acquire conductivity either due to thermal generation of electron-hole pair or due to contribution of conducting electrons by net donor atoms or due to holes contributed by net acceptor atoms.

Metals have partially filled conduction band or overlapping conduction and valence bands. As a result, there are mobile electrons available in copious amount. This amount is of the order 10^{22} per centimeter cube. As we see in the Table(1.10), atomic concentration in Solids are of the order of 10^{22} per c.c. and each atom contributes an electron for conduction. Hence availability of conduction electrons is 12 orders of magnitude greater than intrinsic Silicon and 9 orders of magnitude greater than that of intrinsic Germanium. Because of this large number of mobile carriers present in metal that resistivity is so much lower in metal than that in semiconductor. This point will become clearer as we proceed with the quantum-mechanical interpretation of conducting electron in metal or in semi-conductor.

As seen in Table(1.10) , the mobility of conducting electrons in semiconductors is much higher than that in metal which is typically 44cm^2/(V-sec) for Copper.

Table(1.10) Characteristics of Ge, Si and GaAs at 300 K.

Characteristics	Symbol	Units	Ge	Si	GaAs	Cu
Effective Density of States	**N_c**	**Cm^{-3}**	**1.04×10^{19}**	2.8×10^{19}	**4.7×10^{17}**	
	N_v	**Cm^{-3}**	**6.1×10^{18}**	1.02×10^{19}	**7.0×10^{18}**	
Energy Gap	**E_g**	eV	0.68	1.12	1.42	
Intrinsic Carrier concentration	**n_i**	**Cm^{-3}**	**2.25×10^{13}**	1.5×10^{10}	**1.1×10^{6}**	**8.5×10^{22}**
Effective mass	**m_n (unit mass 9.11×10^{-31} Kg)**		0.33	0.33	0.068	
	m_p (unit		0.31	0.56	0.56	

	mass 9.11 × 10 -31 Kg)					
Mobility	**µ n**	**Cmˆ2/ (V s)**	3900	1350	8600	44
	µ n	**Cmˆ2/ (V s)**	1900	480	250	
Dielectric Constant	**εr**		16.3	11.8	10.9	
Atomic Concentration		**Cmˆ -3**	**4.42 × 10ˆ 22**	**5.0 × 10ˆ 22**	4.42 × 10ˆ22	**8.52 × 10ˆ22**
Breakdown Field	**EBR**	V/cm	**10ˆ 5**	**3 × 10ˆ 5**	**3.5 × 10ˆ 5**	

1.10.2.1. METALS (*with special reference to the mobility of conducting electrons and its implications for particle accelerators*).

Metal is a lattice of positive ions held together by a gas of conducting electrons. The conducting electrons belonging to the conduction band have their wave-functions spread through out the metallic lattice. The average kinetic energy per electron is (3/5)E_F (This will be a tutorial exercise). Hence

$$\frac{1}{2} m^* v_e^2$$

=

$$\frac{3}{5}E_F$$

1.93

where m* is the effective mass of the electron but we will assume it to be the free space mass.

Therefore:

$$v_e = \sqrt{[\frac{6}{5} \times \frac{E_F}{m^*}]}$$

……………. 1.94

Where

$v_e =$ **velocity of the mobile electron in conduction band**

This velocity is not thermal velocity but velocity resulting from Pauli's Exclusion Principle which essentially is the result of the ferm-ionic nature of electrons. Electrons tend to repel one another when confined in a small Cartesian Space. Electrons are claustrophobic.

Therefore mean free path $=$

$$L^* = v_e \times \tau$$

…… 1.95

Where τ is mean free time.

Substituting the appropriate values for each metal, we get the mean free path for electron in their respective metals.

Table(1.10) Tabulation of the Fermi Energy, velocity, mean free time and mean free path of conducting electrons in their respective metals.

Metal	**E F**	**Velocity(x 10ˆ 5 m/s)**	**τ (femtosec)**	L*(A°)
Li	4.7	9.96	9	90
Na	3.1	8.08	31	250
K	2.1	6.65	44	293
Cu	7.0	12.15	27	328
Ag	5.5	10.77	41	441.6

As we see from Table(1.10), the mean distance between two scatterers is 2 orders of magnitude greater than the lattice constant which is of the order of 5 A°. Hence lattice centers per se are not the scatterers but infact the disorderliness is what causes the scattering. *The scatterers are thermal vibrations of lattice centers, the structural defect in crystal growth and the substitutional/interstitial* *impurities.* This implies that with reduction in temperature mobile electrons will experience less scattering hence the metal will exhibit less resistivity leading to positive temperature coefficient of resistance. We will dwell upon this in Section (1.12).

1.10.2.2. SEMICONDUCTORS

Semiconductors are insulators initially. At low temperatures, all electrons are strongly bonded to their host atoms. Only at temperatures above Liquid Nitrogen that thermal generation of electron-hole pairs take place. So in semiconductors the situation is quite different as compared to that in metal. The conducting electrons and holes owe their mobility to thermal energy they possess in contrast to the conducting electrons in metal. On an average by Equipartition Law of Energy, the mobile carriers possess (1/2)kT thermal energy per carrier per degree of freedom. Since the carriers have 3 degrees of freedom hence they possess (3/2)kT average thermal energy per carrier.

Therefore

$$\frac{1}{2}m^* v_e^2 = \frac{3}{2}kT$$

…………………… **1.96**

Therefore

$$v_e = \sqrt{[\frac{3kT}{m^*}]}$$

……………………………………………………………… 1.97

Therefore in semiconductors the mean free path will be the product of the thermal velocity and the mean free time. Mean free time is calculated from mobility of the mobile carriers which is determined experimentally.

From Table(1.10) we obtain the mobility values. In Table (1.11) the mobility, mean free time, thermal velocity and mean free paths are tabulated for Ge , Si and GaAs.

Table(1.11) Mobilities, Mean Free Times, Thermal Velocities and Mean Free Paths of Ge, Si and GaAs.

Semiconductor	**µn (cm^2/(V sec))**	τ (femtosec)	**v e (m/sec)**	L* (A°)
Ge	3900	2217	**0.95×10^5**	2106
Si	1350	767.6	**0.95×10^5**	729

GaAs	8600	4890	**0.95×10ˆ** **5**	4645.5

As we see electron has much larger mobility in semiconductors as compared to that in metals. This implies that the mean free path of electrons is greater by one order of magnitude in semiconductor as compared to that in metal. But why is the scattering less in semiconductors as compared to that in metal.? This answer is obtained by determining the de Broglie wavelength of electron and by using wave optics.

We will determine the velocity of a conducting electron in Electron Microscope, in metal and in semiconductor. In these three cases the conducting electron gains Kinetic Energy equal to the Potential Energy it loses while falling through a potential difference of 10kV in case of Electron Microscope(because 10kV is the accelerating voltage in Electron Microscope), through a potential difference of 4V in case of metal(because average kinetic energy associated with conducting electron is (3/5)E_F and E_F is 7eV in copper) and through a potential difference 0.025V in case of semiconductor (since thermal voltage at 300K Room Temperature is kT/q = 0.025V). From the kinetic velocity the de Broglie wavelength is determined. The set of equations are: Kinetic Energy gained =

$$\frac{p^2}{2\times m^*} = q \times V_{acc}$$

Therefore momentum gained

$$p = \sqrt{2m^* q V_{acc}}$$

;

Therefore de Broglie wavelength:

$$\lambda = \frac{h}{p} = \frac{h}{\sqrt{2m^* q V_{acc}}}$$

;

In Table (1.12) the de Broglie wavelengths are tabulated:

Table1.12. de Broglie wavelengths of conducting electron in Electron Microscope, Metal and Semiconductor.

	V acc	**v e (m/ sec)**	λ(m)	Implications
Electron	10kV	**59×10ˆ 6**	**10ˆ -11 m**	λ< < a

Microscope			**= (1/50) (5A°)**	(lattice constant)
	100kV	**187.6×10 6**	**4×10ˆ -12 m**	
Metal	4V	**10ˆ 6**	**6×10ˆ -10 m = (5A°)**	λ~ a (lattice constant)
Semiconductor	0.025V	**10ˆ 5**	**7.75×10ˆ -9 m = (78A°)**	λ > > a (lattice constant)

As seen from Table(1.12), we see that de Borglie wavelength is much less than the lattice constant in case of Electron Microscope. For 100kV , theoretically the resolution should be (1/100)(4A°) This is like Sunlight falling through a broad aperture. Sun-ray will pass in a straight line and shadow of the aperture should fall on the screen behind the aperture. Hence in an Electron Microscope, a regular lattice array does not scatter an electron beam. The shadow of the crystal lattice should be imaged. But this theoretical resolution is never achieved since we are using magnetostatic focusing. Only 1A° is the resolution actually achieved. In case of 10kV, though the theoretical resolution (1/50)(5A°) but in practice only 10A° resolution is achieved. The electron beam can penetrate through a thin specimen and produce the image of its broad features without being influenced by the atomic details.

In electron microscope the electron can be accelerated to higher energy to obtain a finer resolution. It can resolve on the scale of molecules but can barely perceive the atoms.

To resolve at atomic and sub-atomic level we need to go to particle accelerators. *Particle Accelerators are gargantuan machines which can be regarded as giant microscopes for probing into the innermost recesses of matter - an awesome complement to the giant telescopes which probe to the edges of the Universe* .

To arrive at the resolving power of particle accelerator we must know Special Theory of Relativity and we must make relativistic corrections in order to arrive at the correct resolving power of the particle accelerators. These are described in the Appendix XXXXIV . Here we will just use them to arrive at the resolving power of the particle accelerators.

Relativistic momentum is related to the total energy E by the following relation ship:

$$E^2 = p^2c^2 + m_0^2c^4$$

……………………………………………………………………………

1.98

de Broglie wavelength associated with this particle is:

$$\lambda = \frac{h}{p} = \frac{hc}{\sqrt{[E^2-(m_0c^2)^2]}}$$

……

1.99

Using Equation (1.99), the resolving power of various particle accelerators operational around the world is tabulated in Table(1.13).[Taken from Table(9.1), “Overview of Particle Physics”, by Abdus Salam, New Physics, edited by Paul Davies, Cambridge University Press, 1992]

Table 1.13. The resolution of the particle accelerators around the World.

Name & Location	Energy reached	Year	Resolution	Particle detected
Rutherford Manchester,UK.	Alpha decay10MeV,alpha particle’s velocity = **2 × 107 m/s,**Alpha particle = 4He nucleus;	1911	**4.5 × 10ˆ -15 m**	Nucleus**size** = **10ˆ-14m;**Ruther determined the size to be 30fmBut the

				correct estimate is 7fm;
		1919	**1.24×10^{-15} m**	Protons size = **10^{-15}m = 1fm;**
		1932	**1.24×10^{-15} m**	Neutrons size = 1 **= 1fm**
	1GeV		**1.24×10^{-15} m**	
Bepc(e + e-) Beijing	4GeV	1987		
TRISTAN(e + e-) Japan	60GeV	1987		
	10GeV	1979	**1.24×10^{-16} m**	Quarks **size = 10^{-16}m**
SLC(e + e-) Stanford, California, USA;	100GeV	1987	**1.24×10^{-17} m**	**W-, W+ & Z0** detected
LEP(I) (e + e-) Large electron-positron collidor CERN, Geneva;	100GeV	1987	**1.24×10^{-17} m**	
LEP(II) (e + e-)CERN, Geneva	200GeV	1995	**6.2×10^{-18} m**	Top Quarks detected
HERA(ep) Hamburg	320GeV	1991	**3.87×10^{-18} m**	
SpS CERN, Geneva	900GeV	1986	**1.38×10^{-18} m**	
Tevatron FermiLab, USA	1TeV	1987	**1.24×10ˆ**	No

				-18 m	excited state of quarks or leptons detected **size** = **10^{-18}m**
Tevatron	FermiLab, USA	2TeV	1987	**6.2×10^{-19} m**	
UNK	Serpukhov, Russia	3TeV	1995	**4.13×10^{-19} m**	
Ee	Serpukhov, Russia	4TeV	?	**3×10^{-19} m**	
Large HadronCollider(LHC),CERN,Geneva		16TeV	?	**7.75×10^{-20} m**	
SSC(super particle superconductingCollider),USA;		40TeV	?	**3.1×10^{-20} m**	
		1PeV	?	**1.24×10^{-21} m**	
		1EeV	?	**1.24×10^{-24} m**	

* the first particle accelerator was established at Cavendish Laboratory, Cambridge University. In 1919 Rutherford became the first Director and he was instrumental in establishing the particle accelerator.

In Metal the wavelength is comparable to the lattice constant. This is like light falling through a narrow aperture whose dimension is comparable to the wavelength. Incident light will form a circular diffraction pattern behind the aperture on the target screen. This implies that conducting electron in a metallic lattice is strongly scattered by the lattice centers. Hence it has a very low mobility.

In Semiconductor, the de Broglie wave length is much larger than the lattice constant. Hence lattice scattering is weak and only the gross imperfections cause the scattering. These gross imperfections could be phonons and dislocations extending over several lattice constants. This is what makes conducting electrons much more mobile in semiconductor as compared to that in metal.

In metal, conducting electrons behave like degenerate gas and not quite like ideal gas whereas in semiconductors they behave like non-degenerate gas which is more like ideal gas obeying ideal gas law.

In ideal gas the molecules are far apart, independent of one another and possessing average energy of (3/2)kT whereas in degenerate gas the molecules are closely packed and average kinetic energy is much larger than (3/2)kT. In Table(1.14),

Metals and Semiconductors parameters have been tabulated in the same table.

Table(1.14). Conductivity(σ), Fermi Level(E_F), Mean Free Path(L*) and Mean Free Time(τ) at

0°C for monovalent metals and semiconductors.

Metal	σ(10^6 S/cm)	ρ(Ω-cm)	n(10^22/cm3)	μ(cm^2/(V-s)) = σ/(nq)	E_F (eV)	L*(A°)	τ(fs) = meμ/q × 10^-4
Li	0.12	8.3 × 10^-6	4.62	16.2	4.7	110	9
Na	0.23	4.35 × 10^-6	2.65	54.17	3.1	350	31
K	0.19	5.26 × 10^-6			2.1	370	44
Cu	0.64	1.67 × 10^-6	8.5	47	7.0	420	27
Ag	0.68	1.47 × 10^-6	5.9	72	5.5	570	41
Ge		47	ni = 2.5 × 10^13	3900		2106	2217
Si		300k	ni = 1.5 × 10^10	1350		729	767.6
GaAs		70.5M	ni = 2 × 10^6	8600		4645.5	4890

SSPD_Chapter 1_Part 11_Solid State of Matter_diifferent bondings

Chapter 1_ Part 11 continues with the discussion on different kinds of bonding in Solid State of matter

Chapter 1_Part 11_Solid State of Matter_ Bonding.

1.10.3. BONDING IN SOLIDS

In Group IV elements there is 100% sharing of electrons leading to 100% covalent bond. 4 electrons in outer most orbit of an atom are shared with 4 electrons of the four neighboring atoms. In the process the given atom fulfills its octave condition of eight electrons as well the four neighboring atoms also fulfill their respective octave condition. This leads to the formation of very stable and strong covalent bonds. Covalent bond fully depends upon time sharing of outer most electrons with the neighboring atoms.

In contrast NaCl has ionic bond created by columbic attraction between positively charged Na(Group I) ion by virtue of losing its lone outer most electron to Cl(Group VII) which becomes negatively ionized. This ionic bond has 100% transfer of electron from Na to Cl.

1.10.3.1. COVALENT BOND.

Covalent bond of Group IV elements exhibit a very interesting feature. Though all elements of Group IV have covalent bonds they exhibit varying physical properties. As we move down from Carbon to Silicon to Germanium to Gray Tin and finally to Lead we find hardness, melting point and electrical conductivity changing over a wide range. Diamond , an allotrope of Carbon, is the hardest material and has very high melting point of 3550ºC whereas Tin is very soft and has low melting point. (It has just been announced that “Diamond is no longer nature’s hardest material”,Jessica Griggs, New Scientist, 16th February,2009. Wurtzite Boron Nitride is 18% harder and Mineral Lonesdaleite is 58% harder than diamond.) Diamond is very good insulator, Si and Ge are semi-conductors and Tin is a good conductor. In fact one finds a gradual transition from a extreme valence bond in Diamond to metallic bond in Lead. Diamond, Silicon, Germanium, Silcon Carbide , tin and rutile form covalent bonds.

The gradual transition from extreme valence bond to metallic bond occurs because of gradual reduction in the energy band gap between Valence Band and Conduction Band. Energy Band Gap is 5.2eV in Diamond. Whereas it is 1.21eV in Si, 0.67eV in Ge, 0.08eV in Gray Tin and 0eV in Lead. This leads to increasing intrinsic carrier concentration and hence to increasing metallicity as we move from Diamond to Gray Tin. In Table 1.15 we make a comparative study of Group IV elements.

Table 1.15. Room Temperature Energy Band Gap(E_g), intrinsic carrier concentration (n_i) and resistivity(ρ)

	Diamond(C)	Si	Ge	Sn(gray)	Pb
E_g (eV)	5.2	1.21	0.67	0.08	0
n_i (cm^{-3})	**5.46×10^{-2}**	**1.5×10^{10}**	**1.3×10^{13}**	**5.5×10^{18}**	**2.6×10^{19}**
ρ(Ω cm)	**10^{16}**	300,000	47	**2×10^{-4}**	**2×10^{-5}**
bonding	covalent	covalent	covalent	Semi-metallic	metallic

1.10.3.2 IONIC BOND.

Just as elements of Group IV complete their octave condition (closed valence energy shell) by time sharing of electrons, elements of alkali group (Li, Na, K, Rb, Cs) and elements of Group VII, halogen Group(F, Cl, Br, I) complete their octave condition by exchanging electrons. Alkali metal releases its single electron in its outermost orbit to achieve the octave condition and becomes positively charged Cations. Halogen group acquires electron given up by Alkali metal to achieve the octave condition and becomes negatively charged Anion. By acquiring single electron from the alkali metal it completes 8 electrons in outermost orbit where it had 7 electrons from before. By completing this set of 8 electrons it fulfils the octave condition. These two oppositely charged ions bond together to form 100% ionic bond in alkali halide salts. Examples are NaCl,KF, LiF, MgO, CsCl and ZnS. Since both the ions have their electrons tightly bound hence ionic salts are insulators. Since ions are regularly arranged in solid hence they are crystalline in nature: fcc(face centered cube) in NaCl with coordination number 6 and bcc(body centered cube) in CsCl with coordination number 8. [fcc and bcc will be discussed in next section].

Truly ionic crystals cannot absorb light hence they are transparent to light. Light can pass through without absorption. Most ionic salts are hard owing to the strength of the ionic bonds. They have high melting points because of their hardness. They are brittle because the atoms cannot slip past each other. In contrast metals are ductile and malleable because atoms can slip past each other. Ionic crystals are soluble in polar liquids because of their high dielectric constant. The columbic force of attraction is considerably reduced in polar liquids such as water because of the fact that:

$\varepsilon_0\varepsilon_r E = D$

hence $E = D/(\varepsilon_0\varepsilon_r)$

Water has a dielectric constant of 70 under DC condition. In DC condition polarization is contributed in major part due to the alignment of polar molecules which water molecules are. This is the main reason for its high dielectric constant. Hence when NaCl salt is put in water, because of very weak columbic attraction between Cations and Anions, the two dissociate and they give rise to very good electrolytic solution which is suitable for carrying out electrolysis. At the same time ionic salts are insoluable in covalent liquids such as benzene(C_6H_6) and carbon tetrachloride(CCl_4) which have very low dielectric constant. Hence ionic bonds cannot be broken and ionic salts remain intact in organic solvents. The ionic salts are good IR absorbers because IR resonates with the natural frequency of the molecular ionic salt.

On one extreme we have 100% covalent compounds and covalent elements. Covalent bonds may be non-polar if the centres of charge coincide as is the case in Si and Ge and polar if the two centers are displaced as is the case in Water (H_2O).

In water molecule two Hydrogen atoms share their two electrons with six electrons of one oxygen atom. In the process Oxygen completes its octave and Hydrogen also completes its octave and the system is at Energy Minima and hence in stable state. We say two Hydrogen atoms are covalently bonded to one Oxygen atom. In the process center of negative charge shift towards Oxygen nucleus and water molecule becomes polar.

On the other extreme we have 100% ionic bonds between Gr I and Gr VII where there is 100% transfer of electrons.

In between these two extremes we have intermediate cases where we have a mix of covalent bond and ionic bond. As we move from I-VII to II-VI to III-V the ionic bond is progressively transformed to covalent bond. In these case we have mixtures of two bands. Some electrons are exchanged and some electrons are shared. For example II-VI BaS compound is predominantly ionic and slightly covalent. As we move to III-V InP we have predominance of covalent bond.

1.10.3.3. METALLIC BONDS.

A metal is an assembly of positive ions with their valence shells completed and hence octave completed. This assembly is submerged in a sea of conducting electrons which are the outer most electrons of the atoms but which no longer belong to their host atoms but they belong to the whole crystalline lattice. The outermost lone electron is loosely held because of large orbital radius hence first ionization energy is very low. As seen in Table 1.16 for Li it is 5.4eV and as we go down the Group I the size becomes bigger hence first ionization decreases to 3.9eV for Cesium. After first ionization the metal becomes monovalent cations with very stable electron configuration. Hence second ionization is much larger about 14 times more in case of Lithium. Again because of increasing size the valence electrons are less tightly held as we go down Group I hence second ionization also decreases. It is 6 times greater than First Ionization energy in case of Cesium.

Being the first elements of each period, alkali metals have the largest atomic and ionic radii in their respective periods. As we move within a period, the atomic radius and ionic radius tend to decrease due to increase in the effective nuclear charge. On moving down the group, there is increase in the number of shells and, therefore, atomic and ionic radii increase.

Table 1.16. Electronic configurations and ionization energies of Alkali Metals.

	Li Lithium	Na Sodium	K Potassium	Rb Rubidium	Cs Cesium
Atomic Number	3	11	19	37	55
Atomic	152	186	227	248	264

radius(pm)					
Ionic Radius(pm)	76	102	138	152	167
Electronic configuration	(He)2s1	(Ne)3s1	(Ar)4s1	(Kr)5s1	(Xe)6s1
First Ionization Energy (kJ/ Mole)	520	496	419	403	376
First Ionization Energy(eV)	5.4	5.1	4.3	4.17	3.9
Second Ionization Energy (kJ/ Mole)	7298	4562	3051	2633	2230

Within the crystalline lattice, conducting electrons can freely move around hence alkaline metals are conductors but as they come to the boundary they are faced with surface barrier potential φ_{BS} and the energy required to overcome this surface barrier is known as work function $W_F = q\ \varphi_{BS}$. This sea of conductin electrons belong to the whole lattice and hold together the positively charged ionic cores by what we know as metallic bond.. What is the physical basis of this metallic bond?

In the five alkaline metals, in atomic state the outer most electron has zero kinetic energy and negative potential energy decided by the principal quantum number. The principal quantum numbers are 2, 3, 4, 5 and 6 for Li, Na, K, Rb and Cs. As they come together in sold state. The outermost electron acquire Kinetic Energy which is (3/5)E_F . In Cu Fermi Energy is 9.04eV and average Kinetic Energy is 4.44eV. But sum of Potential Energy and Kinetic Energy is at its minimum . Hence metallic solid configuration is stable. But as we move to Group II and Group III, Kinetic Energy increase due to large number of outer orbital electrons or valence electrons hence metallic bond becomes weaker and eventually it is non-existent. Hence metallic bond exists from Group I to Group II (transition metals). It also exists in Group III namely Al, Ga, In and Ti. But not beyond this.

Because of the sea of conducting electrons in conduction band, metals have high electrical conductivities as well as thermal conductivities.

Bonding energies and melting points are shown in Table 1.17.

Table 1.17. Bonding Energies and Melting Temperatures of selected substances.

[“Rudiments of Material Science” by Pillai and Pillai, Talble 1.A, New Age International Publishers 2005**]**

Bonding Type	Material	Bonding Energy	Melting Point	
		kJ/Mole	eV per atom or ion or molecule	(ºC)
Ionic	NaCl	640	3.3	801
	MgO	1000	5.2	2800
Covalent	Si	450	4.7	1410
	C(diamond)	713	7.4	3600
Metallic	Hg(Gr.IIB)	68	0.7	-39
	Al(Gr.III)	324	3.4	660
	Fe(Gr.IIB)	406	4.2	1538
	W(Gr.IIB)	849	8.8	3410
Van der Waals:Induced dipole	Ar(Nobel Gas)	35	0.08	-189
	Chlorine	51	0.32	-101
Van der Waals:Hydrogen bond	Ammonia	35	0.36	-72
	Water	51	0.52	0.0

Because of orderly arrangement of the positive ionic centers in a spatial lattice network, metals are highly crystalline.

The metallic bonds are non-directional and allow the ionic centers to go past one another therefore metals can bend without fracture. Metals are ductile, malleable and possess metallic luster.

Alloys are physical mixtures of different metals but of similar atomic size. Whereas covalent compound or ionic compounds are chemical compounds. Chemical compostion cannot change but stochiometric coefficient of different metals in an alloy can change.

Hydrogen of Group I is an exception. In gaseous, liquid or solid state at normal atmospheric pressure, hydrogen remains non-conducting. This is because Hydrogen is always in H_2 molecular state hence the two electrons are tightly held to the molecular state of Hydrogen.

Under unusually high pressure in Jupiter and Saturn, hydrogen molecules are brought so close together that electrons probability waves overlap and spread to the whole lattice. Under these conditions liquid Hydrogen becomes metallic and conductive. The electric currents set up in the metallic core of Jupiter and Saturn due to the planet's spin create magnetic fields 20 times stronger than the terrestrial magnetic fields.

Metallic Hydrogen if achieved at normal pressure could be used as super conductors, rocket propellants and in fuel cells. Solid Hydrogen Propellants develop 5 times as much thrust per kg as the present day rocket fuels. This is the reason why ISRO is in process of developing cryogenic engine based on liquid-hydrogen propellant rocket system.

Solid Hydrogen in form of Deutrium and Tritium could become very efficient fuel for fusion reactors.

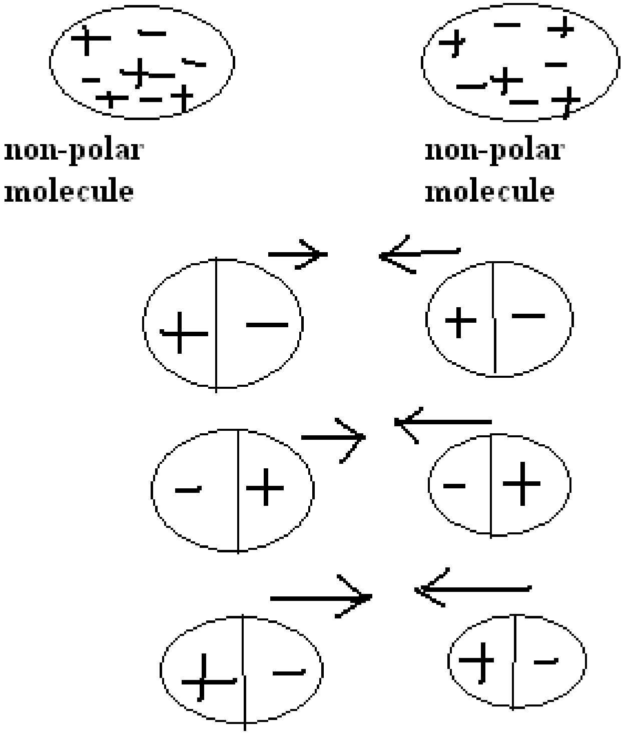
1.10.3.4. VAN DER WAAL’s BOND- INDUCED BONDING .

It has been found over a century that Nobel Gases which are chemically inert and monatomic exhibit inter-atomic attractive forces. It is because of these weak attractive forces that Argon, Helium and Neon liquefy at low temperatures. These forces were first proposed by Dutch Physicist Johannes van de Waals[Appendix XXXXVI] to explain the deviation from the ideal gas laws when applied to real gases. The physical basis of these forces were only recently understood.

In absence of covalent bond, ionic bond and metallic bond, it is Vanm der Waals weak inter-atomic force which provides the cohesive force for liquefaction and solidification particularly in inert elements. The macroscopic behavior of solids such as surface tension, friction, viscosity, adhesion and cohesion have their basis in this force only. This force varies inversely as the seventh power of inter-atomic distance i.e. as $1/r^7$ hence they act only when molecules are very close together.

Van der Waals force act between polar molecules as well as in non-polar molecules. In polar molecules it is simply dipole-dipole electrostatic attraction. In non-polar it is by induction.

The induced bonding is shown in Figure 1.51.



The fluctuations are synchronized so that there is always attractive force on a time average basis

Figure 1.51. The process of induced bonding due to Van der Waals force.

Figure 1.51. The mechanism of generating induced bonding in non-polar molecules.

As shown in Figure 1.51, on a time average basis there may be charge symmetry and generation of dipole but on instantaneous basis the material does not have charge symmetry hence it has no dipole If the dipole generation is properly synchronized then a net force can develop. This is referred to as Van der Waals induced bonding.

Even rare gas atoms experience Van de Waals force which otherwise are chemically inert. It is these forces which help crystallize Argon crystals at -187ºC and liquefy He at 4K.

Van der Waals force is much weaker than ionic, covalent and metallic bonds hence molecular solids , where this fourth bond is operative, have much lower melting point, boiling point and lower bond strengths as shown in Table 1.18.

Table 1.18. Melting Point and bond strength of molecular solids such as Argon, Hydrogen and Methane.

Element or compound	Melting Point(ºC)	Bond Strength (eV/ molecule)
Solid Argon	189	0.08
Solid Hydrogen	259	0.01
Solid Methane	-183	0.1

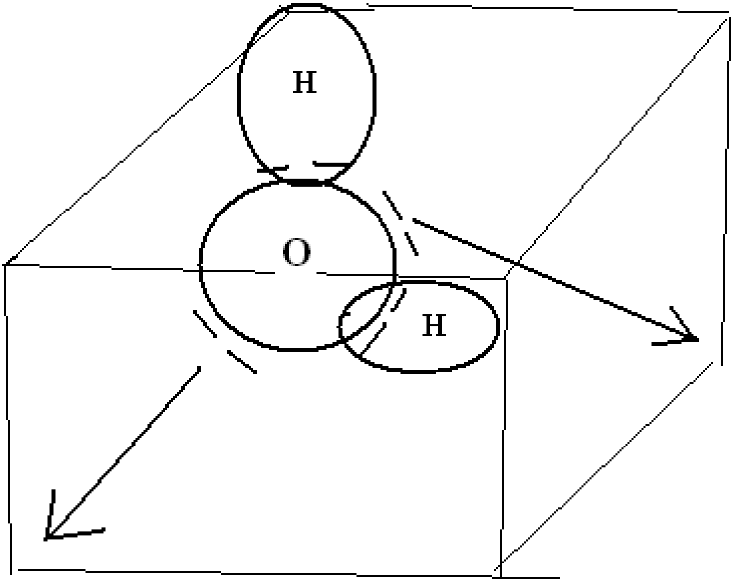
1.10.3.5. VAN DER WAAL’s BOND- HYDROGEN BONDING .

Hydrogen bond is a strong type of Van der Waals bond. This is formed between Hydrogen and Group VII/VI/V/IV elements. Methane (CH_4), Ammonia (NH_3), Water (H_2O) and Hydroflouric Acid (HF) are examples of hydrogen bond based on Van der Waals forces. Here Hydrogen atom or atoms donate their electrons to another element’s atom to complete the octave condition and thereby create stable electronic configuration. The resulting molecule will link up with identical molecules through Van der Waals force. The Van der Waals force is arising due to the electrostatic attraction created by the unshielding of the hydrogen nucleus. This has $1/r^2$ dependence therefore hydrogen bonds are 1/10 th as strong as covalent bonds.

In water molecule, covalent bond exists between one atom of Oxygen and two atoms of Hydrogen. Oxygen completes its octave by sharing its 6 electrons with 2 electrons of 2 Hydrogen atoms. As shown in Figure 1.52. there are 4 pairs of electrons surrounding Oxygen nucleus. These four pairs occupy four regions that form a tetrahedral pattern. Two pairs are linked up with two Hydrogen atoms but two pairs are unattached and are capable of Hydrogen bonding with two other molecules. In next Figure 1.53 we show how H2O link up to form Hydrogen bonds.

In liquid state these bonds are continuous breaking and forming due to thermal agitation. On an average the water molecules are closer together in liquid state than in solid state. In solid state the molecules are permanently arranged in hexagonal channel like structure. Hence water has least density at 0ºC. From 4ºC to 0ºC there is phase transformation from liquid to solid. At 0ºC all the liquid is finally transformed to solid and latent heat of solidification is released. Below 0ºC it remains in solid state. This is known as anolmous behavior of water which permits the upper surface of lakes to freeze keeping the under layer at 4ºC in liquid water state. This permits the marine life to survive in deep frozen weather conditions.

Hydrogen bonds occur in biological molecules. In poly peptide chains of protein where amino acids are connected to form the poly peptide chains, we hydrogen bonds. The two spiral strands of DNA are linked together through hydrogen bonds. Hydrogen bonds provide the optimum strength so as to permit transcription, translation and replication of DNA.



Four pairs of electrons located in tetrahedral pattern. Two lone pairs indicated with arrows can form Hydrogen bond

Figure 1.52. Oxygen-Hydrogen covalent bond illustration in water molecule and the possibility of Hydrogen bonding to two lone pairs of electrons is shown.

Figure 1.52 The covalent bonding between one Oxygen and two Hydrogen atoms and the possibility of Hydrogen bonding with two other H2O molecules indicated by arrows.

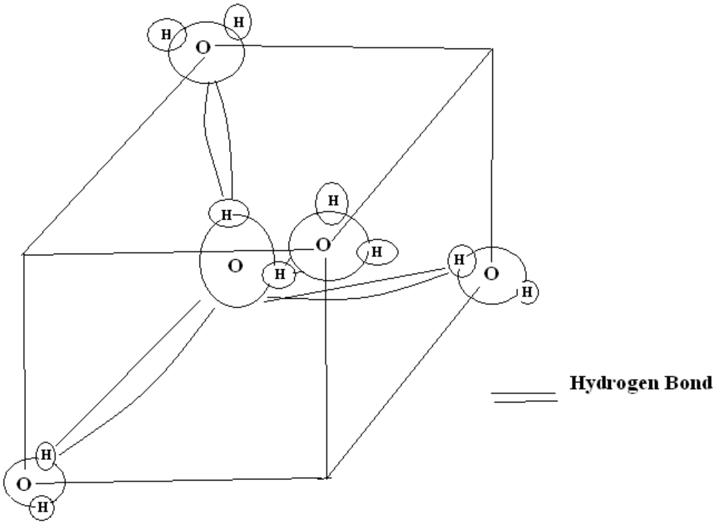


Figure 1.53. The Hydrogen bonding of one H2O molecule with four other H2O molecules.

Table 1.19. Comparison of Bond Types and typical properties.

[“Principles of Electronic Materials and Devices”, Secnd Edition, S. O. Kasap, McGraw Hill, 2002]

Bond Type	Typical Solids	Bond energy (eV/atom)	Melt. Temp. (°C)	Elastic Modulus (GPa)	Density (g cm^{-3})	Typ. Propertie
Ionic	NaCl (**rock salt**); MgO (**magnesia**);	3.2 10	801 2852	40 250	2.17 3.58	NaCl:electrica insulator. Becomes

										conductive at high temperatures. elastic modulus. Hard and brittle but cleavable. Thermal conductivity less than metals
Metallic	**Cu; Mg;**	3.1	1.1	1083	650	120	44	8.96	1.74	**Cu:good electrical conductor.**M thermal conductor. High elastic modulus. Ductile.
Covalent	**Si;C(diamond);**	4	7.4	1410	3550	190	827	2.33	3.52	Si: large elastic modulus. Hard , brittle.C(diam Good

							electricalInsu Exceptionally high thermal conductivity.
Van derWaals: Hydrogenbonding	PVC **(polymer); H20 (ice);**	0.52	2	49;1		1.30.917	PVC: low elastic modulus. Some ductility. Ice: electrical insulator. Poor thermal conductivity. Large thermal expansion coefficient.
Van derWaals: Condensedbonding	Crystalline **Argon;**	0.09	-189	8		1.8	Low elastic modulus. Electrical insulator. Poor thermal conductivity. Large thernmal expansion

coefficient.

SSPD_Chapter 1_Part 11_Solid State of Matter_Summary of Bonds
SSPD_Chapter 1_Part 11_ Solid State of Matter gives a perspective view of different categories of bonds.

SSPD_Chapter 1-Part 11_ Solid State Matter_Summary of Bonds

1.10.4. SUMMARY OF FOUR KINDS OF SOLID STATE BONDS.

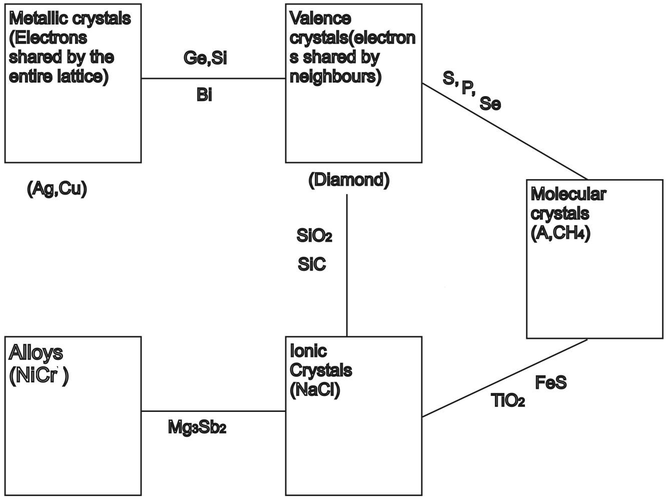

Metallic crystals (Electrons shared by the entire lattice)
(Ag,Cu)
Ge,Si
Bi
Valence crystals(electrons shared by neighbours)
(Diamond)
S, P, Se
Molecular crystals (A,CH4)
SiO2
SiC
Alloys (NiCr)
Ionic Crystals (NaCl)
Mg3Sb2
FeS
TiO2

Figure 1.54. Summary of four kinds of bonds

In Figure 1.54 we have classified solids according to their bonds. The five boxes in the Figure give five distinct classes of solids:

1. Metallic Crystalline Solids-the valence electrons are shared by the entire lattice;

2. Covalent Crystalline Solids- the valence electrons are shared by the immediate neighbours; In between these two categories we have Ge, Si and Bi which are part covalent and part metallic at room temperature.
3. Metallic Alloys which are physical mixtures of two similar atomic sized metals such as NiCr(Nichrome). Here stoichometric coefficient can be arbitrarily fixed;
4. Ionic crystals where electrons are transferred from Group I, II, III to Group VII, VI, V respectively to create ionic pairs extending in all directions through out the lattice.

In between alloys and ionic crystals we have Magnesium Antimonide(Mg_3Sb_2).

This is a definite chemical compound with a definite chemical composition quite unlike the physical mixture of alloy but it is not purely ionic. A case like this is in between alloy and ionic crystals. When two metals differ chemically to an appreciable extent then only such an intermediate condition occurs.

In between ionic crystals, where electrons are transferred between neighbours, and covalent crystals , where electrons are shared on time divison basis, we have SiC(Silicon Carbide) and SiO_2(Silicon Dixide). In these cases we have sharing as well as transfer of electrons for achieving the energy minima configuration.

1. E. Van der Waals crystals where dipole-dipole interaction and higher order interactions are at work such as Methane and Argon. In between Van der Waals crystals and ionic crystals we have FeS and TiO_2. Here layer wise we have ionic bonds and between the layers we have Van der Waals bond.

In between Van de Waals crystals and covalent crystals we have Graphite, Phosphorous, Sulphur and Selenium. Here layer wise we have covalent bond and inter layer bonds are weak Van der Waals bonds.

SSPD_Chapter 1_Part 11_Solid State of Matter_Crystalline Nature of Solids
Chapter 1_Part11_Solid State Matter_ Crystalline Nature describes the 14 Bravais Lattices and three allotropic forms of Carbon.

SSPD_Chapter1_Part 11_Solid State of Matter_Crystalline Nature of Solid.

1.11. CRYSTALLINE STRUCTURE- SINGLE CRYSTAL, POLY CRYSTAL AND AMORPHOUS.

1.11.1. SINGLE CRYSTAL

Solid State is essentially a condensed state of matter. All solids are composed of atoms, molecules or ions. The atoms are held together by ionic bond, covalent bond, van der wal's bond or metallic bond. When a geometric configuration of atoms/molecules/ions is repeated in 3-Dimensions to generate a solid then we have a crystalline solid. The geometric configuration is called the Unit Cell. The Crystal is a periodic arrangement of these unit cells and these unit cells are the basis or the pattern of the crystal. The periodic arrangement of atoms is referred to as crystal lattice and the unit cells are referred to as the lattice centers. The unit cells are of seven basic geometric configurations as listed in Table (1.20). The unit cell have three basic reference vectors a, b, c and mutual angles of inclination α, β, γ as shown in Figure (1.54). Different combinations of three basic reference vectors a, b, c and mutual angles of inclination α, β, γ give rise to 14 different 3-D spatial configurations or space lattices called BRAVAIS LATTICE. These are 1 triclinic, 2 monoclinic, 4 rhombic, 2 tetragonal and 3 cubic. But these reduce to seven basic geometric configurations which are illustrated in Figure 1.55.

Table 1.20. The seven crystal systems.

Serial Number	Name of the Crystal System	Length of axes.	Angle between axes.	Examples
1	Simple Cube	$a = b = c$	$\alpha = \beta = \gamma = 90°$	Cu, NaCl
2	Tetragonal	$a = b \neq c$	$\alpha = \beta = \gamma = 90°$	**SnO_2, $NiSO_4$**
3	Orthorhombic	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90°$	**KNO_3, $BaSO_4$**
4	Monoclinic	$a = b \neq c$	$\alpha = \beta = 90 \neq \gamma$	**Na_2SO_4, $FeSO_4$**
5	Triclinic	$a \neq b \neq c$	$\alpha = \beta = 90° \neq \gamma$	**$CuSO_4$, K_2Cr_2O**
6	Trigonal (Rhombohedral)	$a = b = c$	$\alpha = \beta = \gamma \neq 90°$	**Ca SO_4, As**

7	Hexagonal	$a = b \neq c$	$\alpha = \beta = 90^\circ$ and $\gamma = 120^\circ$	**Zn, Cd, SiO 2**

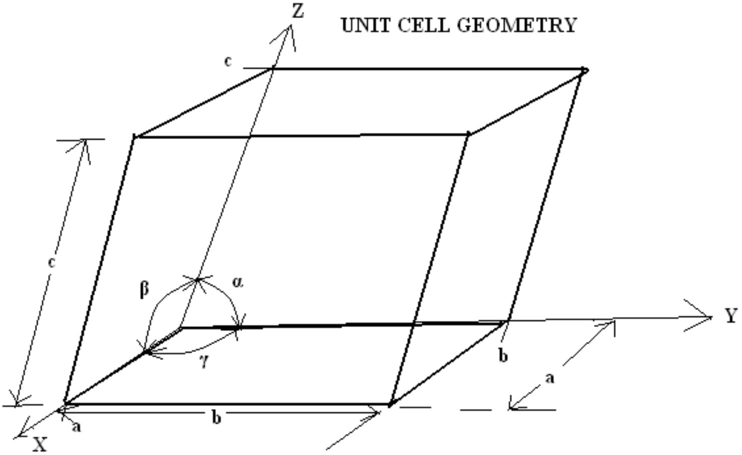


Figure 1.54. A parallelepipid is chosen to describe geometry of an unit cell. X, Y and Z axes lie along the edges of parallelepipid where lower-left corner is the origin of the the reference system.

Orthorhombic or Orthogonal Crystal structure, $a \neq b \neq c$; $\alpha = \beta = \gamma = 90^\circ$; left to right: simple, base centered, body centered, face centered.

Cubic crystal structure, $a = b = c$; $\alpha = \beta = \gamma = 90^\circ$; simple, b.c.c. , f.c.c..

Tetragonal crystal structure, $a = b \neq c$; $\alpha = \beta = \gamma = 90^\circ$; simple and b.c.c.

Out of 14 Bravais Lattices, 9 have unit cells with orthogonal edges.

Two monoclinic lattices(referred to orthogonal planes for clarity) $a = b \neq c$; $\alpha = \beta = 90^\circ$, γ; (a) simple and (b) centered;

Hexagonal Lattice . $a = b \neq c$; $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$;

Triognal Lattice or Rhombohedral. It has a set of primitive vectors of equal lengths and equal angles. $a = b = c$; $\alpha = \beta = \gamma \neq 90^\circ$;

Triclinic Lattice. No 90º angle. For clarity it has been referred to a cube and to orthogonal planes.

The diagrams of all these seven basic geometric configurations will uploaded as a supplement.

Figure. 1.55. The topograghy of the seven crystalline structure and 14 Bravais Lattices.

[Semiconductor Device Electronics, Warner and Grung, Publisher: Holt, Rinehart and Winston,1991, pp.64-65]

Table 1.21. Comparison of cell properties of some crystal lattices.

S.No.	Properties	Diamond cube(dc)	Simple Cube(sc)	Body Centered Cube(bcc)	**Face** Centred Cube(fcc)	Hexa-gonal Close_pack structure	NaCl
1	Vol. of unit cell	a^3	a^3	a^3	a^3	$(3/2)$ **$[\sqrt{3}(ca^2)]$**	a^3
2	No. of atoms percell	8	1	2	4	6	4 cations4 anions
3	No.of atoms per unit vol.	**$8/ a^3$**	**$1/ a^3$**	**$2/ a^3$**	**$4/ a^3$**	$4/$ **$[\sqrt{3}(ca^2)]$**	**$8/a^3$**
4	**No.of nearest neighbours**	4	6	8	12	12	
5	Nearest neighbourDistance(2r)	$a\sqrt{3}/4$	$a/2$	$a\sqrt{3}/2$	$a\sqrt{2}/2$	$a/2$	
6	Atomic radius	$a\sqrt{3}/8$	$a/2$	$a\sqrt{3}/4$	$a\sqrt{2}/4$	$a/2$	
7	Atomic packingFactor	$\pi\sqrt{3}/16 = 0.34$	$\pi/6 = 0.52$	$\pi\sqrt{3}/8 = 0.68$	$\pi\sqrt{2}/6 = 0.74$	$\pi\sqrt{2}/6 = 0.74$	0.67
8	Examples	**Ge, Si and Diamond**	Polonium	Sodium, Iron	Aluminium, Copper, Lead	Magnesium, Zinc, Ca	Ionic solids such as NaCl, AgCl,LiFMgC

Water shows an anomaly in the range from 4°C to

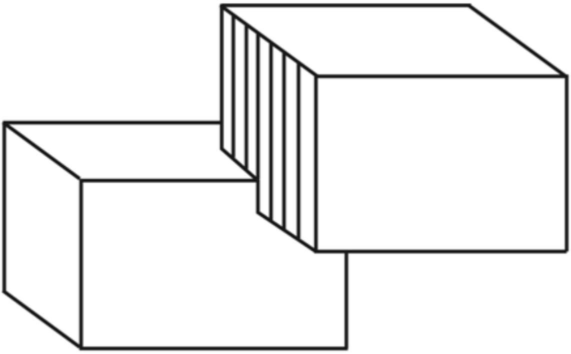
0°C. With reduction in temperature at and below 4°C the density decreases instead of increasing. At 0°C, the Ice Temperature, it has the lightest density. This is because of the open-pack hexagonal structure of ice crystal. From 4°C to 0°C, the phase transition is occurring through out the volume of liquid and at 0°C the phase transition is complete. At this temperature the whole mass of water is fully transformed into a solid piece of ice. Therefore at 0°C we have the lightest density of Ice. It is this anomaly which helps the survival of fishes in frozen lakes. Nine-tenth of ice-berg is submerged and one-tenth of the ice-berg is visible. Therefore we say that the tip of the ice-berg is visible.

1.11.2 ALLOTROPIC FORMS OF CARBON

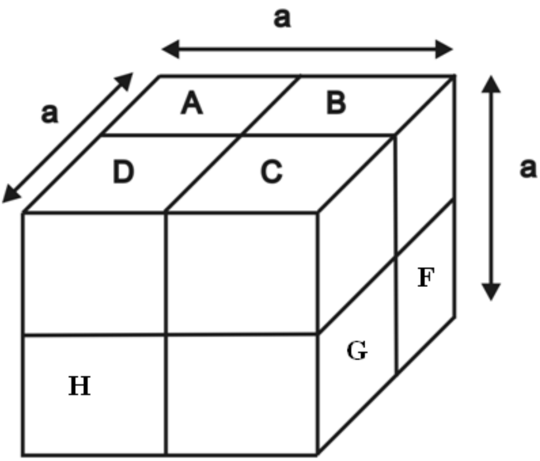
There are elements which exist as different crystalline allotropes such as Carbon. This most commonly found element exists as three crystalline allotropes namely Graphite, Diamond and Buckminstefullerene Crystal also known a buckeyball structure. The structure is shown in Figure 1.56 and Figure 1.57 and their properties are tabulated in Table (1.22).

1.11.2.1. DIAMOND CRYSTAL.

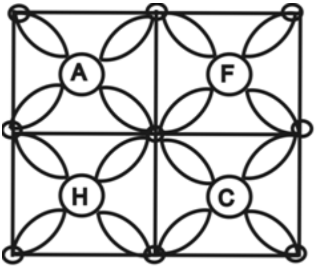
In Figure 1.56 we show the Diamond allotropic form of Carbon. As shown in Fig 1.56 A , Diamond is two interpenetrating FCC sub-lattices with one sub-lattice displaced with respect to the other along the diagonal of the cube by a quarter of diagonal length (a√3)/4 where ‘a’ the lattice parameter. Since Diamond unit cell is a cube, its lattice parameter is the side of a cube as shown in Figure 1.56 B. The basic unit cell of Diamond is cube of equal sides ‘a’ length. ‘a’ is the lattice constant of Diamond. Its value is 3.57 Angstrom. The basic unit cell is subdivided into 8 subcubes ABCDEFGH. ABCD comprise the upper 4 subcubes and EFGH comprise the lower 4 subcubes. In the Figure 1.56B , subcube E is hidden due to the perspective view. The subcube A, C, in the upper layer and F,H in the lower layer contain the tetrahedral structure of carbon atoms as shown in Figure 1.56D. As a result the basic UNIT CELL has 8 C atoms at the 8 corners of the cube, 6 C atoms on the six faces of the cube and 4 C atoms in the bodies of A, C , F, H subcube. The net result is that on looking at the PLAN VIEW of Diamond Unit Cell as shown in Figure 1.56C we see one C atom sharing its 4 valence electrons with 4 C atoms at its 4 corners. This kind of structure is repeated through out.



1.56A Two interpenetrating FCC sub-lattices with one sub-lattice displaced w.r.t the other along the diagonal of the cube by a quarter of diagonal length i.e by $\frac{a\sqrt{3}}{4}$

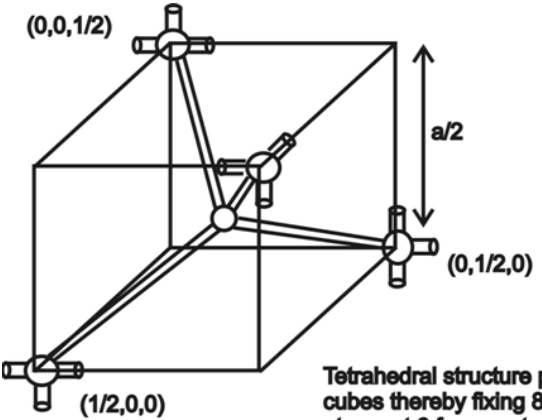


1.56B A major cell divided in 8 sub-cells (A,B,C,D,E,F,G,H) (Perspective view) E subcell is hidden



Each atom is surrounded by 4 neighbouring atoms to complete the octave condition and achieve covalent bonds

1.56C Plan View



Tetrahedral structure present in A,C,F,H sub-cubes thereby fixing 8 atoms at 8 corners, 6 atoms at 6 face centers,4 atoms at 4 body centers of 4 sub-cubes(A,C,F,H)

Figure 1.56D. Geometrical configuration of one subcube.

1.11.2.2._GRAPHITE_AND_GRAPHENE

[‘Graphene: Status and Prospects’, A.K. Gelen, Science, Vol 324, 19th June 2009, 1530-1534;

‘Carbon Woderland’, Andre K. Geim and Philby Kim, Scientific American, April 2008, 90-97]

The two other alltropic forms of Carbon are Graphite and Buckminsterfullerene[Appendix

XXXXV]. The crystal structure of these two allotropes are shown in Figure 1.57. The comparative study of these three alltropic forms of carbon is shown in Table 1.22.

In Graphite each layer has hexagonal arrangement of carbon atoms. As seen in Figure 1.57, each carbon atom has three neighbouring carbon atoms covalently bonded. This arrangement leaves 1 electron unbounded hence free to be released through out the lattice. This makes Graphite a good conductor and it has a metallic luster. There is a weak Van der Waals induced bonding between the layers as already discussed in Bonding chapter. Hence layers can move past one another and flak easily. They find applications in lubricants and pencil lead.

Graphite is a 3D bulk material which consists of stacks of 2D one atom hexagonal rings of Carbon sheets. In recent years with the advent of nanotechnology , it has been possible to separate out individual sheets and experiment with it. The flat sheet of hexagonal rings of Carbon is known as GRAPHENE. Wrapped up graphene is known as fullerenes. When it is wrapped up in honey comb cylinder like structure it is called carbon nano-tubes and when it is wrapped as soccer ball shaped as in Figure 1.57 then it is called bucky ball.

This 1 atom thick 2D grapheme is a wonder material. It is the strongest material- stronger than steel. The orderly arrangement of hexagonal rings is repeated through out the sheet with remarkable regularity as a result even at room temperature of 300K electron's mean free path is in micrometer. This implies very few and far between scattering. Hence under the influence of electric field, electrons behave like mass-less Dirac quasi-particles on which relativistic quantum mechanics or Quantum Electrodynamics applies. Electron travels at (1/300) of speed of light. This is in contrast to electrons in 3D semiconductor where mass is near that in free space and mobility and drift velocity is much lower. These useful properties has opened immense possibilities for grapheme based electronics. Ballistic FETs are emerging which are suitable for 100GHz and above application. Graphene has given a new lease of life to Moore's Law of microminiaturization. It will enable the scaling down right up to 10 nm.

Under normal conditions graphite crystallizes out but under excessive pressure condition diamond is favoured. Once the diamond has crystallized it remains stable under normal temperature and pressure. Deep in Earth's vegetation remains are transformed into diamonds and coal over eons. This is the source of natural diamond.

Synthetic diamond can be obtained in the following manner. Graphite is dissolved in molten cobalt or nickel and the mixture is compressed by 60,000 bar pressure at 1600K. Less than 1 mm wide diamonds can be synthesized in this way. It is widely used in industrial applications for cutting and grinding tools.

1.11.2.3 BUCKY BALLS AND NANOTUBES.

In 1986 at Rice University, Houston, Texas, a third allotrope of carbon was discovered. It was fcc crystal structure as shown in Figure 1.57. At each lattice point there was a giant molecule. Each giant molecule C_{60} consisted of 60 carbon atoms and it was of soccer ball type geometry. Within this giant molecule there are 12 pentagons and 20 hexagons. Since soccer ball type geometry resembles Geodesic Dome designed by Architect R. Buckminster Fuller[Appendix XXXXV] hence in his honour it is called ‘Buckminsterfullerene’ or bucky ball in short. These are stable and chemically unreactive. It can be made from graphite in the laboratory. It is found in small quantities in soot and in carbon rich found in Russia. Fullerene molecules are held in a solid form by Van der Waals attraction just as hexagonal carbon layers of graphene are held in stack in graphite. Fullerene come in groups of 28, 32, 50 70. There can be still larger fullerene molecules. K_3C_{60} yield a superconductor at room temperature.

Carbon Nanotubes is a variant of bucky balls. It contains tiny cylinders of rolled up grapheme. The hexagon of carbon atoms may be arranged straight along the tube or wind around like a helix. In the former case nanotube is a conductor and in the latter case it behaves like a semi-conductor. If nano tubes can be fabricated in longer lengths they hold promise as very strong fibers ten times stronger than stel and six times lighter and flexible. They will act as ideal material for reinforcing epoxy resin giving rise to unusually strong composite mnmaterial.

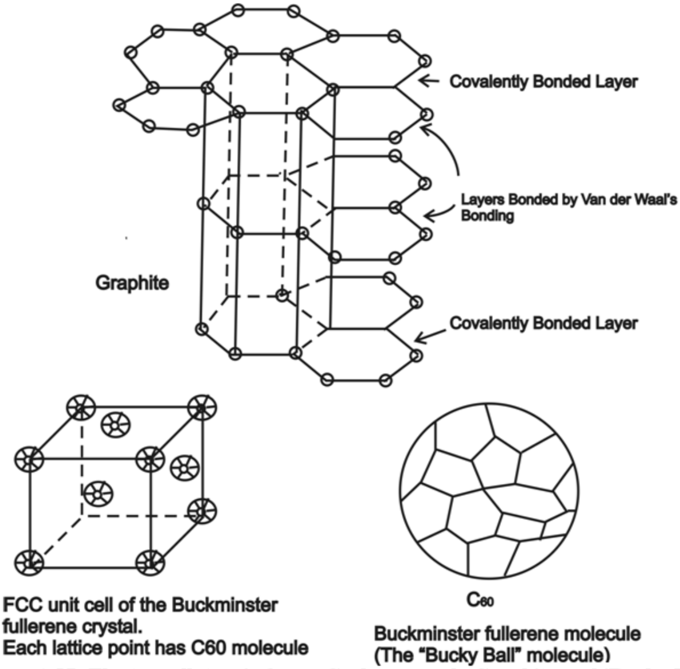

Covalently Bonded Layer
Layers Bonded by Van der Waal's Bonding
Graphite
Covalently Bonded Layer
FCC unit cell of the Buckminster fullerene crystal.
Each lattice point has C60 molecule
C60
Buckminster fullerene molecule
(The "Bucky Ball" molecule)

Figure 1.57. The two allotropic forms Carbon namely Graphite and 'Bucky Ball'.

Table 1.22. Crystalline allotropes of Carbon(ρ is the density and Y is elastic modulus or Young's modulus)

[From *Principles of Electronic Materials and Devices,* *Second Edition,* **S. O. Kasap(© McGraw- Hill, 2002)]**

	Graphite	Diamond	Buckminsterfullere
Structure	Covalent bonding within layers.Van der Waals bonding between layers.Hexagonal unit cell.	Covalent bonded network.Diamond crystal structure.	**Covalently bonded C 60 spheroidal molecules held in an FCC crystal structure by van der Waals bonding.**
Electrical and Thermal Properties	Good electricalConductor.Thermal conductivity comparable to metals.	Very good insulator.**Excellent thermal conductor, about 5 times more than that of silver or copper.**	Semiconductor.**Co with alkali metals (e.g. K3C60) exhibit superconductivity**
Mechanical Properties	Lubricating agent.Machinable.Bulk graphite:$Y \approx$	**The hardest material.**$Y = 827$GPa,ρ	Mechanically soft.$Y \approx 18$GPa$\rho =$

	27GPaρ = **2.25 gm/cm3**	= **3.5 gm/cm3**	**1.65 gm/cm3**
Comment	Stable allotrope at atmospheric pressure.	High pressure allotrope.	Laboratory synthesized. Occurs in the soot of partial combustion.
Uses and potential uses	**Metalurgical crucibles,**Welding electrodes heating elements,Electrical contacts,Refractory applications.	Cutting tool applications.Diamond anvils. Diamond film coated drills, blades, bearings etc. Jewellry. Heat conductor for IC's.Possible thin film semiconductor devices as the charge carrier drift mobilities are large.	Possible future semiconductors or superconductivity applications.

Our main interest is Si, Ge , GaAs and other compound semiconductors used in Photonics. The portion of Periodic Table related to Solid State Technology including Photonics is shown in Table(1.23).

Table.1.23. Portion of the Periodic Table related to Semiconductors & Photonics.

Period	Column II	Column III	Column IV	Column V	Column VI
2		B	C	N	
3	Mg	Al	Si	P	S
4	Zn	Ga	Ge	As	Se
5	Cd	In	Sn	Sb	Te
6	Hg		Pb		

1.11.3. POLYCRYSTAL & AMORPHOUS.(Refer Figure 1.39, Chapter1_Part 10_Electron in single crystal and band theory of solid)

Single Crystals exhibit long range periodicity. Single Crystal Silicon Atoms are 4-fold coordinated in exactly the same manner as Diamond crystal. This 4-fold coordination exists throughout the 3D crystalline lattice except at the surface. At the surface there are incomplete bonds which are known as dangling bonds and which give rise to surface states. These surface states introduce traps or recombination centers. These surface states reduce lifetime and reduce mobility of the carriers. The random capture-release of carriers from the conducting channel of MOS devices by these surface states give rise to flicker noise. By Silicon Dioxide passivation layer, these surface states can be inactivated and the adverse effects are minimized.

Silicon Polycrystals exhibit short range periodicity of the order of micrometers. These are grain sizes. These grains or crystal domains are divided by domain/grain boundaries. Grain boundaries are 2D surfaces where 2D defects are present. Defects are present due to broken bonds or due to incomplete bonds. Each domain has its own direction of periodicity and the direction of the periodicity changes from domain to domain. The basic unit cell remains the same in all domains.

Amorphous silicon has a still shorter range of periodicity or no periodicity. This is an irregular structure and has ill defined energy band gap. Because of irregularities there are a large number of dangling bonds within the bulk of the crystal as shown in Figure (1.35). These dangling bonds introduce traps within the bulk of a-Si (amorphous silicon) and hence severely reduce the lifetime as well as mobility of the mobile carriers. While depositing a-Si if hydrogen (or fluorine) introduced then H-atoms “tie –up” the dangling bonds and these improves the electrical properties of a-Si. Hydrogenated a-Si is denoted as a-Si:H.

The technology of growing amorphous silicon is much cheaper than the technology of growing single crystal silicon therefore amorphous silicon solar voltaic cells are emerging as a more economically viable option as renewable energy resource.

Poly-crystal silicon has an unique position in MOS fabrication. MOS stands for Metal-Oxide-Semiconductor. The use of heavily doped poly-silicon layer in place of aluminum as the metal electrode in MOS has proved to be far superior. This is especially true for thinner gate oxides of the order of 100 to 200 A°.

Doped poly-silicon is also used as diffusion source for creating shallow junctions.

Heavily doped Poly-silicon layer on N-Type silicon crystal gives an ohmic contact. By the conventional route of metallization, if we deposit aluminum layer on N Type crystal we obtain a rectifying contact. To ensure an ohmic contact to N type silicon crystal, we have to go for N+ contact deposition in N type silicon and then depositing an aluminum contact layer. This ensures ohmic contact but it entails an extra deposition. Deposition of heavily doped poly-silicon directly on N type silicon crystal gives ohmic contact. This simplifies the fabrication steps.

Using doped poly-silicon, interconnections can be made. Variously doped poly-silicon can be used to fabricate high value IC resistors.

The range of orderliness in single crystal, poly-crystal and amorphous is depicted in Figure (1.39).

Table (1.24) tabulates the properties of crystalline and amorphous silicon.

Table 1.24. Crystalline and amorphous silicon.

[From *Principles of Electronic Materials and Devices, Second Edition,* **S. O. Kasap(© McGraw- Hill, 2002)]**

	Crystalline Sic Si	Amorphous Sia Si	Hydrogenated a Sia Si:H
Stucture	Diamond cubic	Short range order only.Contains micro-voids and dangling bonds.	Short range order only.Structure contains typically 10% H.Hydrogen

			atoms passivate dangling bonds and relieve strain from bonds.
Preparation method	Czochralski technique/ Float zone technique/ epitaxial growth(CVD or Liquid Phase Epitaxy)	Electron beam evaporation of Si.	Chemical vapor deposition(CVD) of silane gas by RF induction heating.
Density**(gm/ cm3)**	2.33	3 to 10% less dense	1 to 3% less dense.
Electronic applications	Discrete and integrated electronic devices.	None	Solar cells,Thin Film Transistor Arrays in Flat Panel Displays.Photocond drums used in photocopying.

SSPD_Chapter 1_Part 11_Solid State of Matter_continued_Direct and indirect band-gap materials

SSPD_Chapter1_Part 11 is a continuation of Solid State of Matter. This describes indirect band gap material and its non-radiative transition charactistics and direct band gap material and its radiative transition characteristics.

SSPD_Chapter 1_Part 11_Continued_Direct and indirect Band gap Semiconductors.

1.11.4. DIRECT AND INDIRECT BANDGAP SEMICONDUCTORS.

Silicon and Germanium are indirect band-gap materials as shown in Figure(1.58.a and b). Whereas compound semiconductors are direct band-gap materials, as shown in Figure (1.59.), with the exception of GaP which need special doping to become direct band gap. When the top of the valence band and bottom of the conduction band occurs at $k=0$ then we have Direct Band Gap material. Here direct transition occurs hence it is radiative transition . Photons can mediate direct transitions. But when the bottom of the conduction band occurs at $k= \pi/a$ and bottom of the valence band occurs at $k=0$ then indirect transition occurs. The extra momentum $\hbar k = \hbar.(\pi/a)= h/(2a)$ cannot be removed by photon. Phonon has larger momentum and phonon mediates an indirect transition. This leads to increase in thermal vibration of crystal lattice. Phonon is a quanta of lattice vibrational energy. Hence indirect transition is non-radiative. If the edge of conduction band is made of s-orbital electrons then we have direct band-gap material. If the edge of conduction band is made of p-orbital then we have indirect band-gap material.

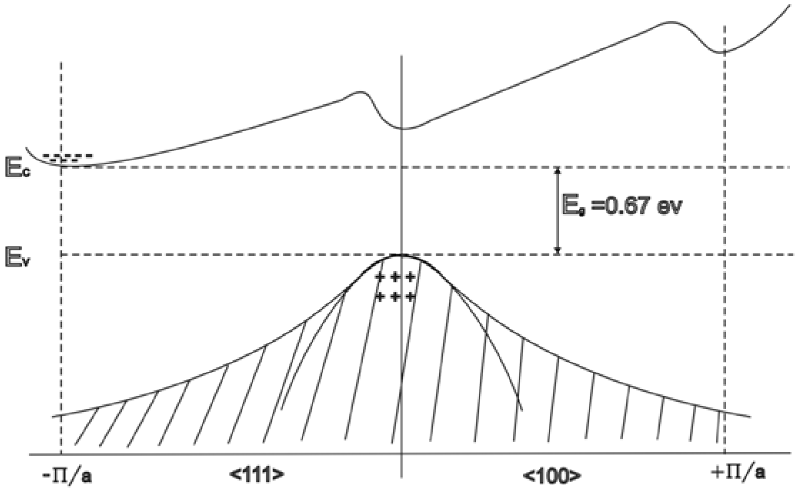


Figure 1.58. (A) Germanium

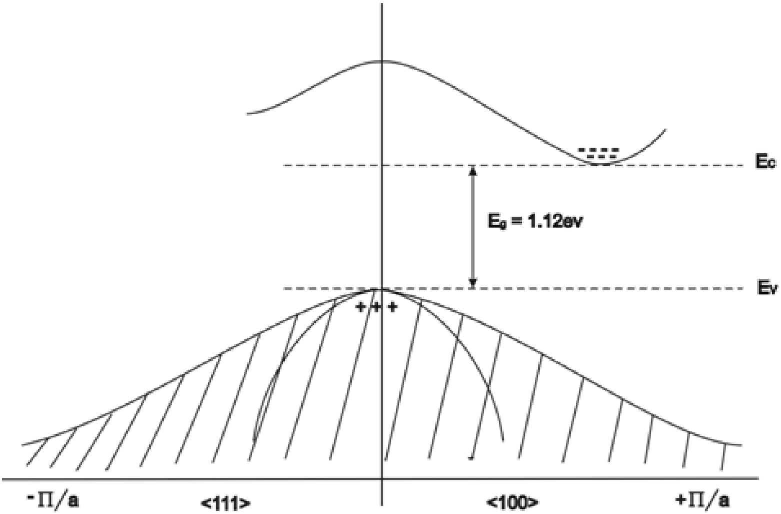
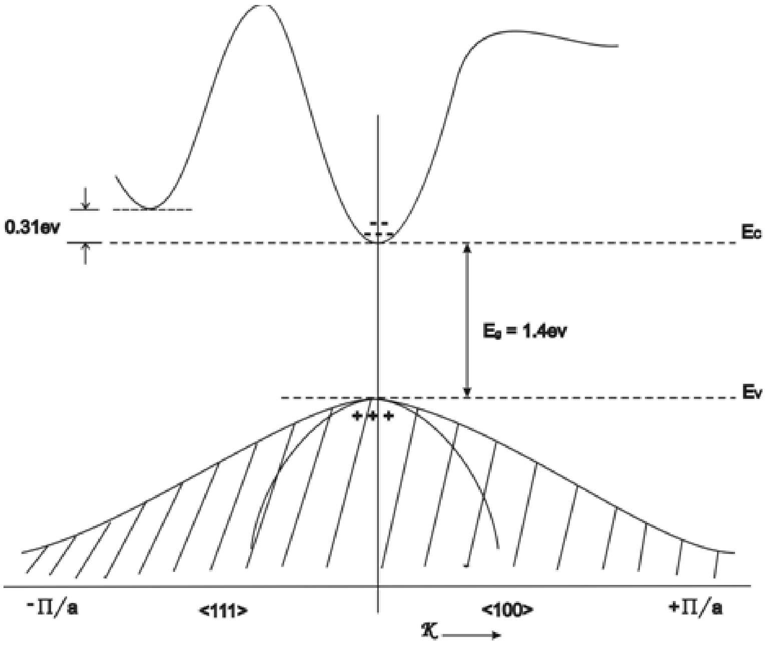


Figure1.58 (b) Silicon



(C) GaAs,InSb,CdS

Figure 1.59. E-k diagram of GaAs

Figure.1.58 and 59. E-k diagram of Ge, Si and GaAs

1.11.4.1. PHONONS.

A quanta of lattice vibrational energy is known as phonon. A Crystal Lattice System can be treated as lattice centers interconnected with bonds equivalent to elastic springs. These springs, representing the atomic bonds, behave like harmonic oscillators which has quantized energies $E_n = (n+1/2)\hbar\omega$. At 0K, the Zero Point Energy of the lattice is $(1/2)\hbar\omega$. The Lattice vibrations propogate like coupled vibrations or as cooperative vibrations of many atoms. Lattice vibrations are longitudinal (atomic vibrations are in the same direction as that of propogation) as well as transverse (atomic vibrations are at right angle to the direction of propogation) as shown in Figure (1.60). These lattice vibrations travel as lattice waves. They occur at 4×10^{12} Hz and at 14×10^{12} Hz .At 4×10^{12} Hz the phonons are known as Transverse Acoustical (TA) Waves and as Longitudinal Acoustical (LA) Waves. At 14×10^{12} Hz, these are known as Transverse Optical (TO) Waves and as Longitudinal Optical (LO) Waves.

The velocity of propogation $= \sqrt{(B/\rho)}$

where B is the elastic bulk modulus and ρ is the density of the crystal.

Substituting the appropriate values B= 70Gpa and

$\rho = 2.7$ gm-cm^{-3} for Aluminium:

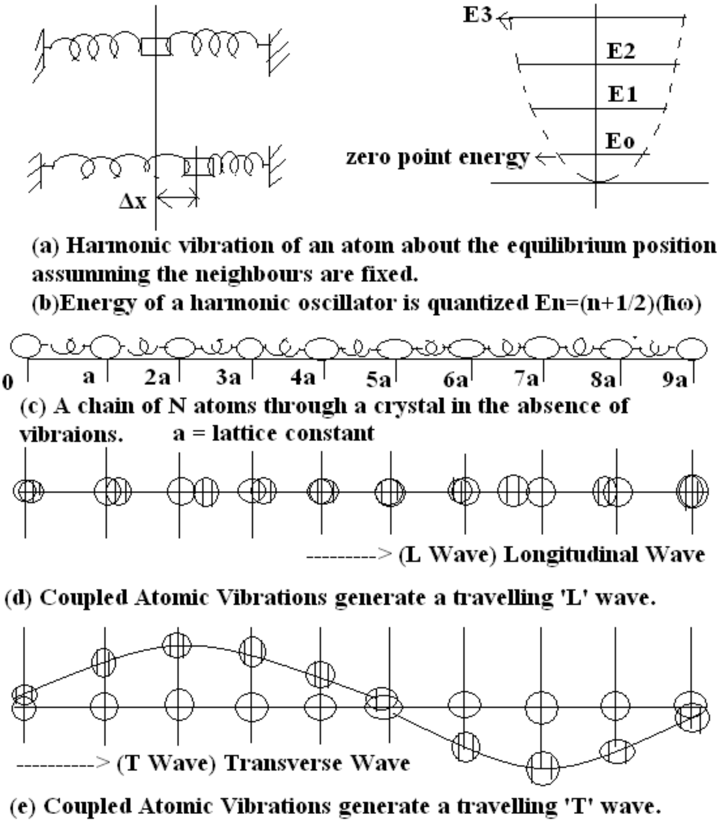
Velocity of propogation is 5092m/s in Aluminum.

The frequency of oscillation is $\sim 10^{13}$ Hz.

Therefore λ_{phonon} = velocity/ frequency = 5 A°;

Linear momentum of phonons = **pphonons = h/ λ = 10-24 Kg-m/sec** = ħk;

Therefore **kphonons = 10-24 Kg-m/sec/ħ** = **2**π×**1010 radians/m**;


E3
E2
E1
Eo
zero point energy
Δx
0
a
2a
3a
4a
5a
6a
7a
8a
9a
---------> (L Wave) Longitudinal Wave
----------> (T Wave) Transverse Wave

(a) Harmonic vibration of an atom about the equilibrium position assumning the neighbours are fixed.
(b)Energy of a harmonic oscillator is quantized En=(n+1/2)(ħω)

(c) A chain of N atoms through a crystal in the absence of vibraions. a = lattice constant

(d) Coupled Atomic Vibrations generate a travelling 'L' wave.

(e) Coupled Atomic Vibrations generate a travelling 'T' wave.

Figure 1.60. Crystal Lattice System represented as Coupled Harmonic Oscillators.

1.11.4.2 PHOTONS.

Photon is a quanta of optical energy. If we examine Green colored photons of wavelength λ = 500nm

then **pphotons = h/λ = 13.26×10-28 Kg-m/sec**;

We see that $p_{photons} << p_{phonons}$;

Also **$k_{photons} = 2\pi \times 10^{7}$ radians/m**;

In a indirect band-gap material, at the bottom of the conduction band, conducting electrons have $k = \pi/a = \pi/(5 \times 10^{-10})$ therefore $k = (\pi/5) \times 10^{10}$ radians/m whereas holes have $k = 0$. Therefore in in-direct transition the mediating particle must take away the energy as well as momentum. Conservation of momentum can be achieved only through the mediation of phonons ($k_{phonons} = 2\pi \times 10^{10}$ radians/m) because of significant excess momentum present during in-direct transition. Hence in-direct transition is accompanied with the emission of phonons rather than photons. Therefore in-direct transition is non-radiative whereas direct transition is radiative. Therefore Direct Band Gap materials are suitable for light sources.

1.11.4.3. RADIATIVE AND NON-RADIATIVE TRANSITION.

Indirect Transition(momentum as well as energy has to be accounted for):

conducting electron + hole = phonons (non-radiative);

Direct Transition (momentum is zero for both carriers and only energy has to be accounted for :

conducting electron + hole = photons (radiative);

SSPD_Chapter 1_Part 11_Solid State of Matter_continued_Crystal Structure of Si,Ge and GaAs.

SSPD_Chapter 1_Part 11_continuation describes the crystal structure and its crystalline parameters.

SSPD_Chapter 1_ Part 11_Solid State of Matter continued_Silicon, Germanium and GaAs crystal structures

1.11.5. SILICON AND COMPOUND SEMICONDUCTORS.

Before the advent of Solid State Technology, i.e. before 1950, semiconductor was used as two terminal point contact diode and photodiode. With the invention of Bipolar Junction Transistor, Germanium rose to the position of eminence because of its unusually high mobility of electron = 3900 cm^2/(V-sec) [Table 1.10]. But with the invention of Integrated Circuit Chips it yielded ground to Silicon.

Today Silicon has become the staple of the Post-Industrial Society just as Iron became the staple of the Industrial Society.

Why Silicon has become the material of choice and the most studied semiconductor in Computer Era? It is primarily because of Silicon-dioxide layer which can easily be formed by oxidation either in dry

Oxygen or in steam. Oxidation is the key process in I.C.Technology. SiO_2 acts as passivating layer as well as a shielding layer during diffusion. As we know that Silicon is a 3-D orderly arrangement of Silicon atoms which is abruptly terminated at the surface boundary of the crystal. This abrupt termination of orderly arrangement leaves many covalent bonds incomplete. These incomplete bonds are known as dangling bonds. These dangling bonds cause surface states and these surface states tend to trap or release mobile carriers. This random release or trapping of mobile carriers cause 1/f noise or flicker noise. Passivation of the surface states minimizes the flicker noise, stabilizes the electrical characteristics of the device and provides a protective layer to the underlying device. Since SiO_2 is a perfect insulator it is used for isolation of devices which considerably reduces the parasitic effects. Conventionally reversed biased diode is used which eats away a lot of the real estate on the Silicon Surface and gives rise to severe parasitic effects. The real estate is saved as well as parasitics are minimized by the use of SiO_2 as isolation layer as well as the substrate as it is done in Silicon - on - Insulator (SOI) Technology.

The second reason for the popularity of Silicon is easy and cheap availability. It is obtained from silica and silica comprises 25% of Earth’s crust. Silicon is second most abundant after oxygen.

But Silicon is most unsuitable for photonics because of indirect band-gap and its accompanying non-radiative transitions though in recent years Silicon has been suitably modified to obtain photonic circuits on Silicon Substrate. Because of the unsuitability of Silicon, we have had to opt for Compound Semiconductors and its alloy as shown in Table(1.25).

Table(1.25) give the compound semiconductors which have application in Photonics.

Table. 1.25. Compound Semiconductors for Photonics application.

Element	IV-IV compounds	III-V compounds	II-VI compounds	IV-VI compounds
Si	SiC	AlAs	CdS	PbSe.
Ge		AlSb	CdSe	
		BN	CdTe	
		GaAs	ZnS	
		GaP	ZnSe	
		GaSb	ZnTe	
		InAs		
		InP		
		InSb		

In 1980’s with the development of low attenuation optical fiber , short haul-heavy traffic and long haul communication turned towards Optical Fiber Communication. In communication there is a heavy premium on Band Width and Optical fiber could offer immense amount of BW at very low cost. This gave a tremendous boost to R&D in optoelectronic devices and optoelectronic chips. The material of choice could only be Compound Semiconductors and their alloys. The compound semiconductors are direct band-gap materials offering high radiative quantum efficiency.

Indium Antimonide (InSb) was the first III-V compound semiconductor to be discovered in 1950. Because of narrow band-gap, $E_g = 0.17eV$, InSb is the suitable material for fabricating optoelectronic devices in far infrared spectrum because 0.17eV band-gap corresponds to critical wavelength $\lambda_c = (1.24/ E_g) = 7.3\mu m$. This critical wavelength will decide the optical emission and absorption wavelength.

The invention of semiconductor laser and the discovery of Gunn-Effect brought GaAs ($E_g = 1.43eV$) and InP ($E_g = 1.35eV$) to a position of eminence. The need for visible light Light Emitting Diodes by necessity led to the importance of GaP($E_g = 2.1$ eV) and its alloys. 2.1eV band-gap corresponds to a critical wavelength of 0.59μm which falls in visible part of the spectrum. By suitable alloying , any critical wave length in the range 0.4μm to 0.7 μm can be achieved and 0.4μm to 0.7 μm is the visible part of the spectrum. The alloying process, to achieve a suitable critical wavelength, is known as **BAND-GAP ENGINEERING.** We will deal with this topic in PHOTONICS Chapter.

The choice of different Binary Compounds provides us a wide range of Band-gaps ranging from InSb of band-gap 0.17eV to SiC of band-gap 2.99eV. SiC enabled the fabrication of Blue Laser diodes which are used to sense the pixels of DVD (Digital Versatile Disc). The advent of DVD has greatly enhanced the storage capacity. CD stores 700MB whereas DVD stores 4GB and more.

Here it may also be pointed out that Extreme Ultra-Violet photolithography is enabling the fabrication of nano-electronics chips or realization of lateral feature size less than 100nm as required in Ultra-Large-Scale-Integrated (ULSI) Chips. This again is dependent upon research in very wide band-gap material typically a band-gap of 12.4eV corresponding to a critical wavelength of 0.1μm.

The ease of alloying Binary Compounds into Ternary Compounds such as AlGaAs and Quaternary Compound such as InGaAlAs makes the use of compound semiconductor in photonics very attractive and convenient.. The alloyed Binary compounds gives a continuous and monotonic variation in Band-gap energy according the stoichometric coefficient of the alloyed material. This continuous variation gives the Photonic Scientists a more varied option in terms of band-structure, electronic properties and optical properties. The alloying process can be achieved by Molecular Beam Epitaxy(MBE). This alloying process also enables the fabrication of Hetrostructure Junctions which is much more power efficient as compared to Homojunctions. Homojunctions are junctions formed by P type and N Type of same material whereas Hetrojunctions are formed by P Type and N Type of different materials. Table(1.26) tabulates the electronic properties of Compound Semiconductors.

Table.1.26.Properties of Elemental & Binary Compound Semiconductors at 300K.

Semiconductor	Lattice Constant a A°)	Band Gap (Eg eV)	**Band a**	**Mobility b (cm 2 /V-s)**		Dielectric Constant
Elemental				μn	μp	εr
Ge	5.64	0.66	I	3900	1900	16.0
Si	5.43	1.12	I	1450	450	11.9
IV-IV						

SiC[c]	3.08	2.99	I	400	50	10.0
III V						
AlSb	6.13	1.58	I	200	420	14.4
GaAs	5.63	1.42	D	8500	400	13.1
GaP	5.45	2.26	I	110	75	11.1
GaSb	6.09	0.72	D	5000	850	15.7
InAs	6.05	0.36	D	33000	460	14.6
InP	5.86	1.35	D	4600	150	12.4
InSb	6.47	0.17	D	80000	1250	17.7
II VI						
CdS	5.83	2.42	D	340	50	5.4
CdTe	6.48	1.56	D	1050	100	10.2
ZnO	4.58	3.35	D	200	180	9.0
ZnS	5.42	3.68	D	165	5	5.2
IV VI						
PbS	5.93	0.41	I	600	700	17.0
PbTe	6.46	0.31	I	6000	4000	30.0

[a] I = Indirect, D = Direct;

[b] The values are for drift mobilities obtained in the purest and the most perfect materials available to date;

[c] Silicon Carbide crystallizes in the wurtzite structure.(Physics of Semiconductor Devices, S.M.Sze, 2nd Edition, Wiley, New York, 1981.

1.11.6. BASIC CRYSTAL STRUCTURES.

The elemental semiconductors, Ge and Si, crystallize as diamond structure and most of the compound semiconductors crystallize as Zinc-Blende structure. These are variants of cubic structures.

We have three basic cubic structures: simple cube, body centered cube (bcc) and face centered cube (fcc). The side of a cube ‘a’ is known as lattice constant. The three parameters that characterize a cubic unit cell are : coordination number, the packing fraction and the nearest neighbor distance.

Polonium crystallizes as simple cube.

- Each lattice center is occupied by one atom.
- At each corner we have one atom and each corner atom is occupied by eight cubes. Hence unit cell has 8 corner atoms × (1/8) = 1 atom.
- Coordination number is the number of nearest neighbor lattice site.
- In simple cube it is 6. Four lattice sites lie in the same plane and two lie above and below the plane.
- Packing fraction is $\{[(1/8)(4/3)\pi R^3] \times 8\}/a^3$. Here a^3 is the volume of an unit cell. If we assume that the atom spheres are touching one another then $R = a/2$ where R is the radius of the atom. Therefore packing fraction is $\pi/6$. The nearest neighbor distance is ‘a/2’.

Sodium and Tungsten crystallize as Body Centered Cube(bcc).

- This has one extra atom at the body center.
- This amounts to [8 corner atoms × (1/8) + 1] = 2 atoms per cell.
- The coordination number is 8 as the nearest neighbor atoms are the 8 body center atoms residing in the 8 unit cells surrounding each corner atom.
- Therefore nearest neighbor distance $a\sqrt{3}/2$.
- If we assume that the body center atom is touching the corner atom then the radius is $R = a\sqrt{3}/4$.
- Therefore packing factor = (volume of 2 atoms)/a^3 = $[2\times(4/3)\pi R^3]/a^3$ = $[2\times(4/3)\pi(a\sqrt{3}/4)^3]/a^3 = \pi\sqrt{3}/8 = 0.68$.

Aluminum, copper, gold and platinum crystallize as Face Centered Cube(fcc) .

- Here there are no body centered atom. Instead there are 6 atoms at the center of the 6 faces of the cube.
- This amounts to [8 corner atoms × (1/8) + 6 face atoms × (1/2)] = 4 atoms per unit cell.
- The corner atom lies at the intersection of 3 orthogonal planes. Each plane has four faces and each face has one atom at the center which is the nearest neighbor. Hence the coordination number is 12. Nearest neighbor distance is $a\sqrt{2}/2$ and the radius $R = a\sqrt{2}/4$. The packing factor is = (volume of 4 atoms)/a^3 = $[4\times(4/3)\pi R^3]/a^3$ = $[4\times(4/3)\pi(a\sqrt{2}/4)^3]/a^3$

$= \pi\sqrt{2}/6 = 0.74.$

Figure (1.55) in SSPD_Chapter 1_Part 11_Soild state of Matter_Crystalline Nature of Solid depicts the three cubic crystal systems: simple, bcc and fcc.

Silicon and Germanium are Diamond lattice structures. Diamond structure has been described in SSPD_Chapter 1_Part 11_Soild state of Matter_Crystalline Nature of Solid in Figure 56.

- Diamond lattice structure belongs to cubic-crystal family.
- Diamond structure is two interpenetrating fcc sub-lattices with one sub-lattice displaced with respect to the other along the diagonal of the cube by one quarter of the diagonal length i.e. by $a\sqrt{3}/4$ as shown in Figure(1.56.A).
- We can better visualize the diamond structure by dividing one major cube $(a \times a \times a)$ into 8 sub- cubes $(a/2 \times a/2 \times a/2)$ by the name A,B,C,D in the upper half of the major cell and E,F,G,H in the lower half of the major cell as shown in Figure(1.56.B). Subcell E is hidden from sight.
- The perspective view of the major cell is shown in Figure (1.56.B)
- The plan view is shown in Figure (1.56.C) .
- Each major cell has 4 subcells A,C,F,H containing tetrahedral structures of silicon or germanium atoms.

- A tetrahedral structure is shown in Figure(1.56.D). A tetrahedral structure lies in a subcube.
- There is one atom at the center of the subcube.
- This body center atom time-shares its 4 valence electrons with 4 valence electrons of the corner atoms of the subcube.
- The 4 corner atoms forming the covalent bonds with the body centered atom are the two diagonally opposite corner atoms on the upper plane and two diagonally opposite corner atoms on the lower plane but this time it is the other diagonal.
- That is the two sets of diagonally opposite corner atoms comprise two orthogonal diagonals.
- Thus in plan view we see, as in Figure (1.56.C), each Si(or Ge) time sharing its 4 valence electrons with 4 valence electrons of the four corner neighboring atoms and in the process each central atom is fulfilling its octave condition.
- This is 100% covalent bond as the center of positive charge and the center of negative charge are coincident.
- Each atom in the bulk is sharing the 4 valence electrons with 4 neighboring corner atoms hence octave condition and thereby covalent bond is fulfilled every where except at the surface of the crystal.
- Hence at the surface we have unfulfilled

covalent bonds or dangling bonds.

These dangling bonds are responsible for the surface states which if not properly controlled and passivated will lead to serious flicker noise problem as well as to the failure of the devices. At the initial stages of MOS fabrication there was serious Na impurity problem coupled with surface states problem which was leading to very low yields. Andy Grove, who was the cofounder of INTEL along with Gordon Moore and Robert Noyce , came into grip of this problem and gave a permanent solution of the low yields which enabled MOS Technology to take off. MOS Technology is the key to the fabrication of microprocessors.

GaAs has zincblende lattice which is identical to diamond lattice except that one fcc sublattice consists of Ga and the other fcc sublattice consists of As. So in Plan View we see Ga atom surounded by four As atoms and similarly As atom is surrounded by four Ga atoms thereby each atom is fulfilling the octave condition by time sharing the electrons from the 4 neighboring atoms. The plan view of GaAs tetrahedral structure is shown in Figure(1.61).

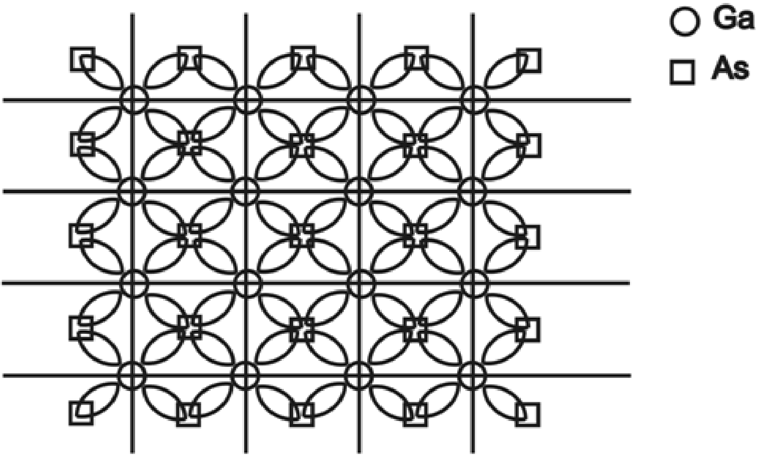


Fig.1.61 . Zincblende structure of GaAs

Figure 1.61. Zincblende structure of GaAs.

1.11.7. CRYSTAL ORIENTATION.

Crystals are anisotropicbecause of different pattern of arrangement of atoms in different directions. Hence etching rate of the crystal is different in different directions. This property is utilized in Integrated Circuit Technology. Therefore crystal orientation is important in device fabrication and hence needs to be clearly defined.

The basis in which we define the planes and directions of a crystalline lattice is made possible by Miller Indices. This is derived from the Cartesian Coordinates in the following manner:

1. Any atom in the crystalline lattice is taken as the origin;

2. Reference coordinate axes are set up in the direction of basis vectors;
3. Consider a system of parallel planes. Find the intercepts made by one of the planes. Express them as multiples of unit basis vectors;
4. Take their reciprocals and reduce them to the smallest triad of integers h,k and l having the same ratio.(hkl) is the Miller Index of that system of parallel planes and <hkl> is the direction normal to the plane;
5. Suppose a plane has intercepts 2a, 3b and 4c where a, b and c are unit basis vectors along X, Y and Z axes. Then reciprocal is (1/2,1/3,1/4). These reciprocals are reduced to the smallest triad having the same ratio. The Miller index is (6,4,3). This has been obtained by multiplying the reciprocals with lowest common multiple of the denominator terms. The denominator terms are 2,3,4. The LCM is 12. Miller Index has been obtained by multiplying the reciprocals by 12.
6. (6,4,3) is the Miller Index of the given plane and <6,4,3> is the orientation vector of the plane.

The normal to the plane along which the crystals cleave is the crystal cleavage plane.. Generally they cleave along <1,1,1> plane. The cleavage plane is the plane of highest atomic density. If the surface of the Si wafer, known as major flat, is parallel to <111> plane then the crystal orientation is <111>. Different planes of a crystal structure is shown in Figure (1.62).

If the surface of the wafer is parallel to YZ plane then the crystal orientation is <100> . For MOS fabrication crystals with orientation <100> is utilized. For other applications <111> crystal orientations are preferred.

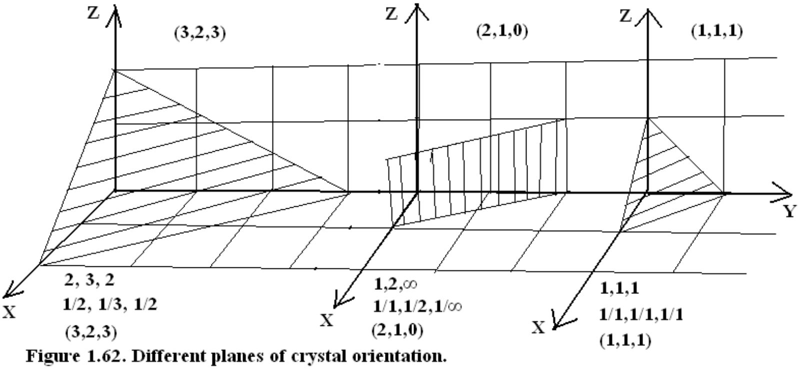


Figure 1.62. Different planes of crystal orientation.

Figure. 1.62. Different Planes of a crystal.

SSPD_Chapter 1_Part 11_Solid State of Matter_Conclusin_Defects in Crystal Structures
SSPD_Chapter 1_Part 11_ Conclusion, describes the 1D,2D and 3D defects in crystals.

SSPD_Chapter 1_Part 11_Conclusion_Defects in Crystal structure

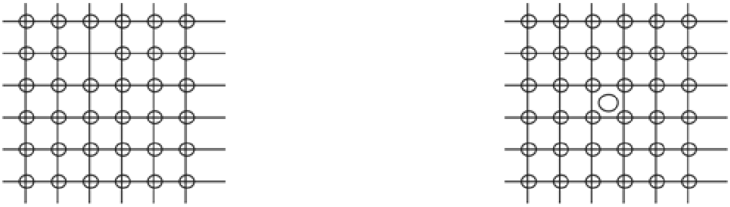
1.11.8. CRYSTAL DEFECTS- 1-D DEFECT, 2-D DEFECTS AND 3-D DEFECTS.

A regular array of atoms extending to infinity in all three dimensions is an ideal crystal. In real life , crystals are always surface bounded where we have dangling bonds and hence surface states. Therefore real crystals are never ideal. But apart from this non-ideality, we have externally introduced defects. This could be a missing atom, atoms out of place, irregular arrangement of atoms, irregularity in the spacing of the rows of atoms and the presence of impurities. These defects effect the electrical properties of the material and devices made out of these materials.

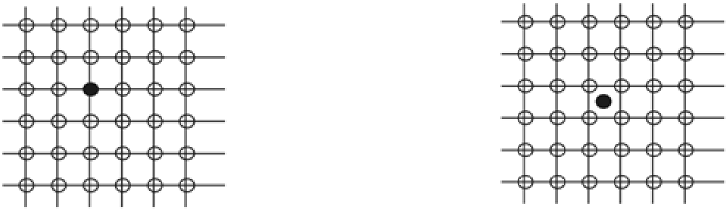
Point defects are the simplest category of defects. Vacancies, interstitial impurity atoms and substitutional impurity atoms comprise point defects. They are shown in Figure (1.63). Vacancies and interstitials occur in all crystals due to thermal excitation of 1~2 eV. As the ambient temperature increases the density of such point defects increases.

By particle radiation such defects can be purposely introduced in a given sample for specific purposes such as decreasing the lifetime of the minority carriers. The lifetime of minority carriers can be changed from millisecond to pico-seconds. A relatively pure, single and defect-free crystal exhibits millisecond lifetime whereas with introduction of defects it falls. Particle radiation induced defects can kill the lifetime to as low as pico-seconds.

The controlled introduction of impurity dopents influence drift mobility, diffusion coefficient and lifetime. The dopents occupy either interstitial positions or substitutional positions.



(a) Vacancy Defect **(b)Interstitial Donor Dopent**



(c) Substitutional Acceptor Dopent **(d) Interstitial Acceptor Dopent**

Figure 1.63. Point defects in a crystal.

Figure 1.63. Point defects in a crystal.(a)Vacancy, (b) Interstitial Donor Dopent (c) Substitutional Acceptor Dopent, (d) Interstitial Acceptor Dopent;

Line Defect(1-D defects): Dislocations are line defects. These act as precipitate sites for metallic impurities. This degrades the performance of the device.

In edge defect, a line of atoms is not in proper position. This causes a permanent deformation in the crystal structure without the loss of a property called ductility. Metals have the maximum ductility because of the nature of metallic bond. We will discuss the nature of metallic bond in the next section. Edge Defect is illustrated in Figure 1.64.

In screw dislocation, a part of the crystal lattice is displaced with respect to the other part. The consequence of this displacement is that atomic layers spiral around this dislocation. Screw Defect is illustrated in Figure 1.65.

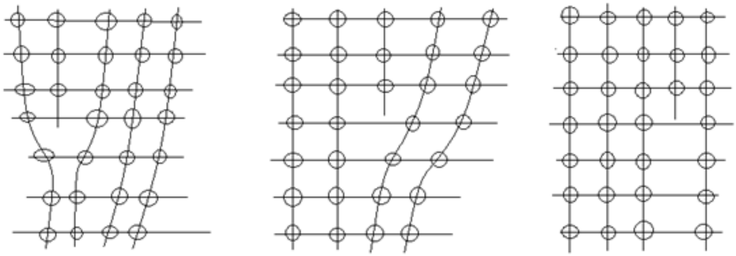

(a) Initial configuration of a crystal with edge dislocation; (b) Edge dislocation moves to right; (c) The crystal has a permanent deformation;

Figure 1.64. Edge Dislocation.

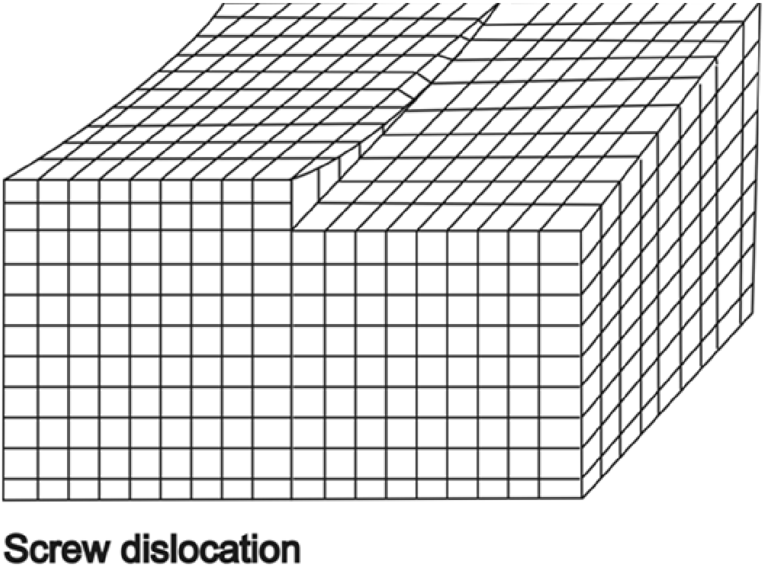


Figure 1.65. Screw Dislocation.

Plane Defects(2-D defects): Solidification of a polycrystalline solid from the melt creates grain boundaries. When crystallographic plane of orientation abruptly changes there we have grain boundaries. Grain Boundaries are shown in Figure 1.66. Grain boundaries lower the mobility of the carriers. Anti-phase boundaries is 2D defect. This occurs in ordered alloy. Here the crystallographic plane direction is the same the layers are in opposite phase. Instead of layer order of ABABABAB…… it is arranged as ABABBABABAB……….Stack faults are also 2-D defect. They occur in a number of crystal structure. But they commonly occur in closed packed structure. Stacking fault is one or two layer interruption in the stacking sequence. Instead of

ABCABCABCABC…….. arrangement as found in fcc structure we have ABCABABABC….. arrangements then we say that stacking fault has occurred.

Volume defects(3-D defects): Precipitates and voids are examples. Precipitates are reduced solubility of dopents at reduced temperature. Voids are shown in Figure 1.67

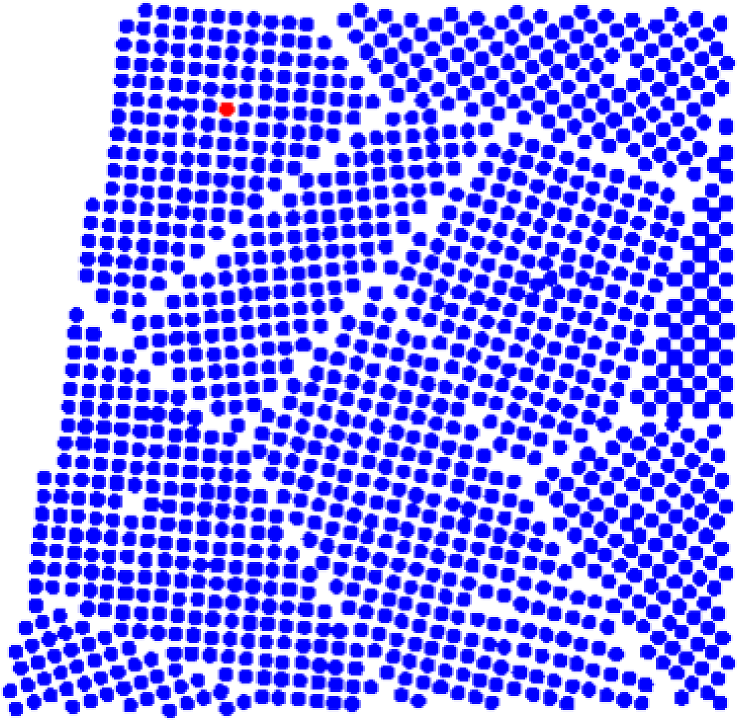


Figure 1.66. Grain Boundaries along with point defects and line defects.

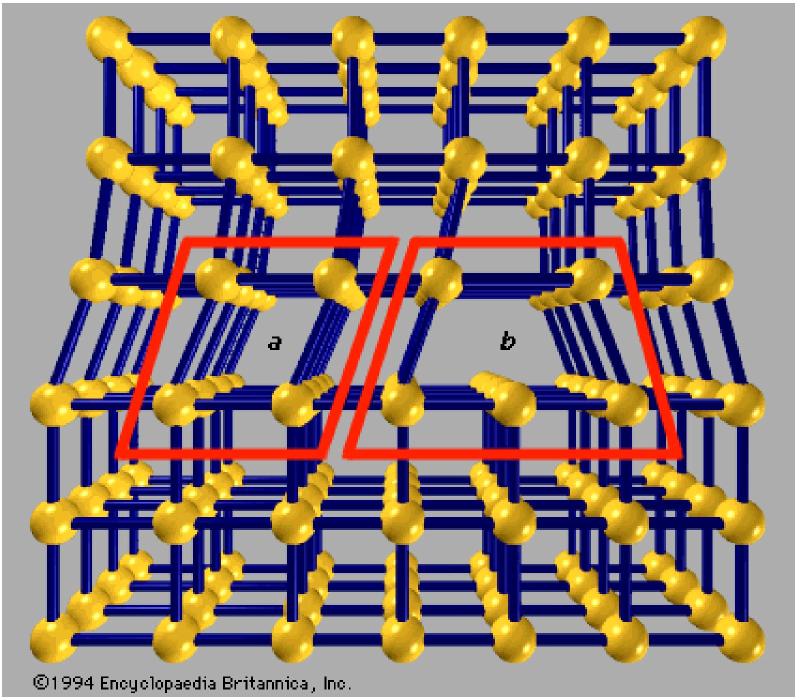


Figure 1.67. Volume Defect such as void.

SSPD_Chapter 1_Part 12_Quantum Mechanical Interpretation of Resistance

SSPD_Chapter 1_Part 12 gives the quantum mechanical interpretation of Resistance in conductors.It is not the lattice parameter but the mean free path which decides the resistance.

SSPD_Chapter 1_Part 12_ Quantum Mechanical Interpretation of Resistance in a conductor.

1.12. SCATTERING OF CONDUCTING ELECTRONS AND RESISTIVITY OF SOLIDS.

Here we are dealing with a 3-D orderly configuration of atoms arranged as a single crystal of infinite dimensions. We have already seen that electron has all the properties of light wave namely: Bragg reflection, refraction , interference and diffraction. Therefore scattering of electrons cannot be understood in a classical manner.

Liquid state of matter causes fifty times more scattering of electron as compared to the gaseous state of matter where as liquid is thousand times more dense. Therefore extent of scattering does not depend on the density of scattering centers but on the disorderly arrangement of the scattering centers. Here it will be proper to point out as to exactly what the difference is between scattering and reflection.

A smooth plane reflects light whereas a rough plane scatters light. In reflection the angles of incidence are identical hence angles of reflection are identical. As a result the total incident beam of light is reflected. In a rough plane, the constituent rays of the light beam are reflected in different directions since the angles of incidence are different for different rays. This is known as scattered light. These two phenomenon are shown in Figure(1.68).

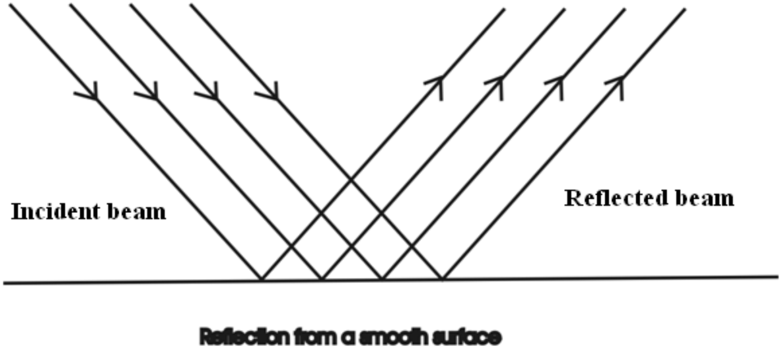


Figure 1.68.a. Reflection from a smooth surface keeps the reflected rays parallel. Hence incident beam is parallel and reflected beam is parallel.

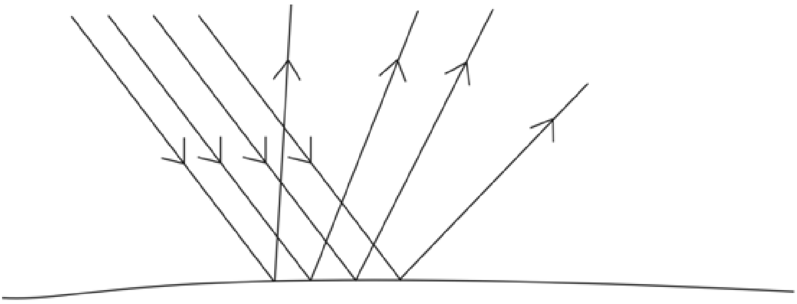


Figure 1.68.b. Parallel beam incident on an irregular and rough surface causes scattering. The reflected rays are no more collimated. They scatter out in all directions. Hence the reflected beam is diffused in nature.

Figure 1.68. Reflection and Scattering.

Electron is a matter wave probability amplitude matter wave. In 3-D orderly crystalline solid the propagation of electron causes the interaction with and vibration of all the intervening lattice centers. The lattice centers become the emitters of secondary wavelets.

The secondary wavelets interfere and propagate forward in the direction of constructive interference.

1. If the crystal is at 0 Kelvin then there is no thermal vibration. Also assume that there are no crystal defects and that there are no impurities.
2. Also the crystalline lattice in not bounded by surfaces and is infinite in the three directions.
3. Also electrons are in the outer partially filled conduction band well removed from the upper edge of the conduction band.

If all these conditions are fulfilled then the direction of constructive interference will be the direction of incidence and electron will propagate forward as if there is no obstacle in its path. The first impulsive energy will sustain the electron in a straight line propagation path with uniform velocity with no dissipation of energy. This propagation will continue for infinite time and till infinite distance. This is exactly as Newton had predicted about the inertness of bodies: A body at rest will continue to be at rest and a body in motion will continue to be in motion in a straight line with an uniform velocity unless made to act otherwise by the application of force.

This is what is more popularly known as SUPERCONDUCTIVITY.

Near the upper edge of the conduction band, electrons will suffer Bragg Reflection and this will be decided by the direction of orientation of the lattice plane from where the specular reflection takes place.

Lattice thermal vibration or/and lattice defects or/ and lattice impurities cause the disorderliness and this disorderliness leads to change of direction of propagation or the change of direction of constructive interference. This change of direction is called scattering of electron. The change of direction is random and the distance over which this happens is statistically varying . The mean distance over which a straight line motion is maintained before scattering occurs is called the mean free path and the mean time taken to cover the mean free path is called mean-free time. The electron scattering within a real crystalline lattice which suffers from all defects and imperfections and which is at Room Temperature has been illustrated in Figure(1.69).

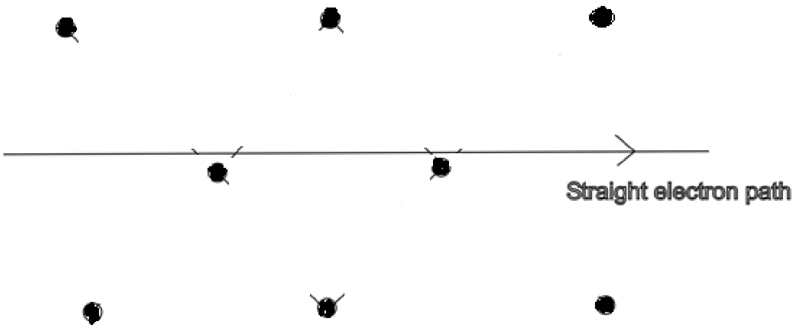


Figure 1.69.a. Electron is moving along a straight path in an ideal crystalline lattice. An ideal crystalline lattice has no defects, no imperfections and no thermal vibratins. Electron experiencesa no scattering.

Black Circles are Lattice Centers arranged in a perfect orderly pattern.

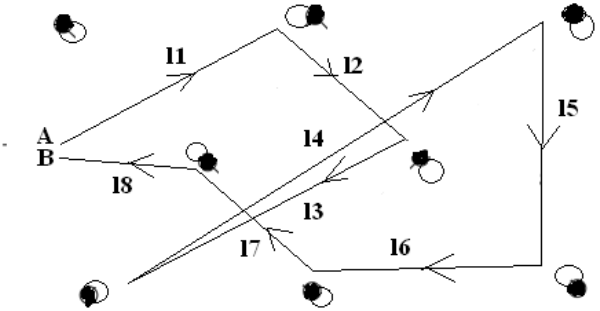


Figure 1.69.b. A real crystalline lattice is shown with lattice centers randomly displaced with respect to orderly arrangement of lattice centers due to theermal vibration. This perturbation in the orderly arrangement leads to random motion of electron with no net displacement in time. As we see that after eight scattering electron initially starting from 'A' finally ends at 'B' which is is more or less in the same location as 'A'.

Black circles denote the orderly pattern of crystal.
Open circle reprent the perturbation in the regular orderly arrangemt of the latice centers.

Figure 1.69. Scattering Phenomena of electron in a real crystalline lattice where lattice thermal vibration, lattice defect, lattice impurities and lattice

boundaries are present.

This scattering of electron does not depend on the density of lattice centers but on the degree of disorderliness of the lattice centers and this degree of disorderliness depends on thermal vibration, lattice defects and lattice impurity density.

By increasing the temperature, the amplitude of lattice thermal vibrations is increased ;

During the growth of the crystal, lattice defects are increased;

During doping and diffusion, impurities are introduced in the crystal.

All these factors contribute towards disorderliness which in turn contribute towards the scattering of the conduction electrons. It is this scattering which creates resistance in a conducting solid. There are two parameters which describe the scattering phenomena:

Mean free path($<l>$) and mean free time(τ).

As seen in Figure(1.69.b), at l_1, l_2, l_3, l_4, ……..distances the direction changes.

Therefore mean free path:

$<\mathbf{l}> = [\Sigma\, \mathbf{l}\; \mathbf{n}\,] /$

N.. 1.100

If the time taken is t_1, t_2, t_3, t_4, …. along the randomly varying paths then the mean free time:

τ = **[Σt_n]**/ **N... 1.101**

And <l>/τ = average thermal velocity= $v_{thermal}$ 1.102

By Equipartition Law of Energy, every degree of freedom has an average thermal energy = (1/2)kT. Since conducting electron has three degrees of freedom(x, y, z) hence the total average thermal energy of a conducting electron is equal to (3/2)kT.

Therefore (1/2)$m_e{}^*$.$v_{thermal}{}^2$ = (3/2)kT 1.103

Where k= Boltzman's constant=1.38×10^{-23}J/Kelvin = 8.62×10^{-5}eV/Kelvin

And T is temperature at Kelvin scale. m_{e^*} = effective mass of electron.

Temperature in Celsius scale is added to 273 to obtain temperature in Kelvin Scale. The zero of Kelvin scale occurs at -273°Celsius. At this temperature i.e. at 0 Kelvin the amplitude of thermal vibrations of lattice centers is zero and if there is no lattice defect and no doping then we have a perfect orderly lattice which will behave like a superconductor.

1.12.1. Quantum Mechanical basis of Resistance in a conducting Solid.

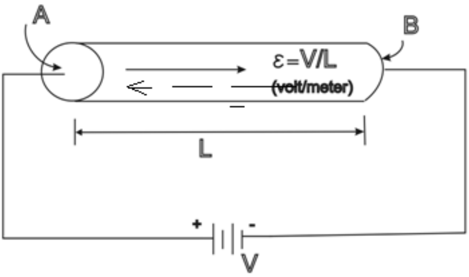
An ideal crystal with no lattice defects, no dopents and at zero Kelvin temperature behaves like a superconductor.

In a normal metal, resistance is always present.

Let us consider a cylindrical metal of length L cm and A(cm)$_2$ cross-sectional area. A potential difference of V volts is applied across it as shown in Figure(1.70).

The electric field along the longitudinal axis is:

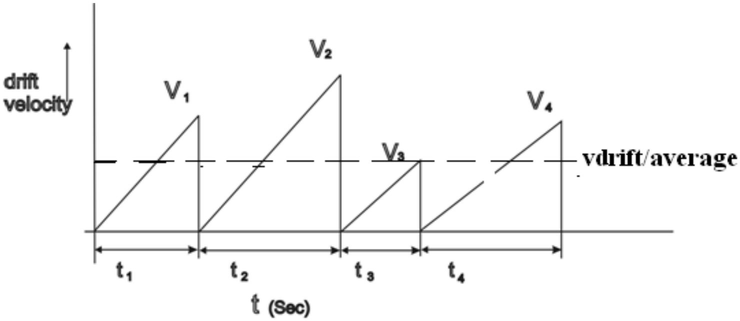
ε = **(V/L) Volt/meter**
……………………………………….. 1.104



(a) The application of an electric field across cylindrical metal specimen

Figure 1.70.a. An electric field is applied along the positive z axis of a cylinderical metal conductor. Electrons move along the negative z axis causing an electric current I = V / R where R = (resistivity).(L)/A=R = ρL/A according to Ohm's Law.

Elecrons move along the negative z axis shown by the broken line.



(b) The instantaneous drift velocity of electron with respect to time

Figure 1.70.b. Instantaneous drift velocity profile with respect to time.

Figure 1.70. Electron drift in z direction(longitudinal axis of the cylinder) after the application of the z-axis electric field.

According to Kinematics, the electric field is applied in z direction. This will cause an acceleration in [-z] direction since electron is negative. This acceleration will continue until the electron gets

scattered. At the point of scattering all the kinetic energy of the electron is imparted to the lattice and the electron starts anew from zero velocity. Since this is an statistical phenomena hence acceleration time periods are t_1, t_2, t_3, t_4,…… and the terminal velocities are v_{t1}, v_{t2}, v_{t3}, v_{t4},……………
Acceleration time periods

t_1, t_2, t_3, t_4,…… are the same as the transit periods along the randomly changing straight line segments as defined in Eq.(1.100)

From kinematics: $v_{tn} = u + at_n$

where a = acceleration = $F/m_e^* = q\varepsilon/m_e^* = q(V/L)/m_e^*$ ………………………………… 1.105

But u = initial velocity = 0 therefore $v_{tn} = at_n = (q\varepsilon/m_e^*)(t_n)$……………………………… 1.106

Therefore the average drift velocity over nth acceleration period is:

$v_n = (v_{tn} + 0)/2 = (q\varepsilon/2m_e^*)(t_n)$ ……………………………… 1.107

Suppose during the flow of current from one end to the other, a given electron undergoes N scattering. This means it undergoes N acceleration periods.

The average drift velocity during N periods of acceleration which occur during the flow of the current is:

$v_{drift} = [\Sigma v_n]/N = [\Sigma(q\varepsilon/ 2m_e^*) (t_n)]/N$

Therefore $v_{drift} = (q\varepsilon/ 2m_e^*)[\Sigma(t_n)]/N$

$v_{drift} = (q\varepsilon/ 2m_e^*)\tau = (q\tau / 2m_e^*).\varepsilon$ …………………………………. 1.108

where $\tau = [\Sigma(t_n)]/N =$ mean free time as defined in Eq.(1.101) and

μ_n = electron mobility $= (q\tau / 2m_e^*)$

This relation was used to determine the electron mobility in Table 1.11, Part 11.

Now under low drift velocity condition, drift velocity varies directly as the applied field ε and the constant of proportionality is known as the drift mobility(μ ($cm^2/(v\text{-}sec)$) as shown in Figure 1.72.

$v_{drift} = (q\tau/ 2m_e^*)\varepsilon = \mu_n \varepsilon$ ………………………………………… 1.109

Current density $= J =$ number of Coulombs/second

Therefore $J = -q \times$ number of electrons flowing through unit cross-sectional area per second.

If we assume that there is no diffusion of mobile carriers and we only have electric drift of the mobile carriers then number of electrons flowing through unit cross-sectional area per second $=(1\text{cm}_2)(v_{drift})$ (n)

where n= number of conducting electrons per unit $(\text{cm})_2$

and v_{drift} = drift velocity of electrons;

Therefore J= (q.v drift .n) Coulombs/(sec-cm 2) ………………………………….. 1.110

Or J = q.μ.ε.n =σ.ε………………………………………………………….. 1.111

Where σ(conductivity) = 1/ρ(resistivity)

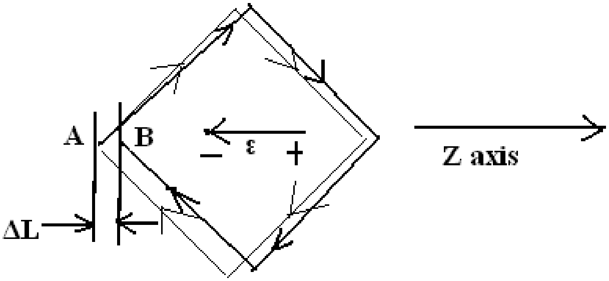
J = q.μ.ε.n =σ.ε

Or J = I/A = (1/ρ)×(V/L)

Therefore V/I = R = (ρL)/A→ Ohm’s Law. ………………………………….. 1.112

In Eq.(1.109) , we find that drift mobility is directly dependent on the mean free time between two consecutive scatterings. Mean free time (τ) is dependent on scattering. Larger is the perturbation in the lattice network from the ideal lattice more frequent will be the scattering and hence shorter will be the mean free time. As the perturbation from the ideal condition or the disorderliness is reduced, so will the scattering phenomena be reduced and mean free time will become longer. Under ideal condition there will be no scattering and mean free time will become infinite. Here the mobility becomes infinite , conductivity becomes infinite and resistivity becomes zero. This is a superconductor. Here once an electron gets an impulsive push it continues to travel in a straight line for infinite distance and for infinite time without any energy dissipation. Hence initial kinetic energy imparted by the impulsive push is conserved forever by the conducting electron. This is tantamount to a current flow in a close loop superconductor without any battery connected to the circuit.

In a normal conductor even with no battery connected, the mobile carriers are undergoing random motion with no net displacement. When an electric field is applied then superimposed on this random motion there is a net displacement of electrons in the opposite direction to the electric field. This has been shown in Figure(1.71). The net displacement is given by ΔL in time Δt.



Thin line is the random motion with no electric field hence there no net displacement.
If an electric field is applied in -z direction then the scattering follows the bolder line and after four scatterings the electron is displaced with respect to the original position by by ΔL in time Δt in positive z direction and the electric field is applied in -z direction.

Figure 1.71. The net drift experienced by an electron under the influence of electric field.

As can be seen in Figure 1.71 electron is undergoing continuous random motion under the influence of thermal energy. As temperature increases electrons become more restless and start meandering along more zig zag path but they never make a net displacement in any direction as shown by ‘thin line’ zig zag path. But as an electric field(ε) is applied in – z direction, electron follows bolder line and as seen in Figure 1.71 it experiences a net displacement of ΔL in +z direction in Δt.

$\Delta L/\Delta t$ = average drift velocity = $v_{drift} = \mu_n \varepsilon$;

In 1911, Kamerling Ons detected superconductivity and superfluidity in solid mercury at 4.3Kelvin.

In 1962 a Russian Scientist was awarded the Nobel Prize in Physics for his study on superfluidity and superconductivity in liquid Helium at 4Kelvin.

In 1987 Karl Alex Muller and Bednorz of Germany were awarded the Nobel Prize for discovering the superconductivity in ceramic Yettrium Barium Copper Oxide [$Y_1Ba_2(CuO)_3$] at liquid Nitrogen temperature 77Kelvin. Bednorz was a research student under Muller at that time.

Before we leave this Chapter on mobility and resistance it will be appropriate to introduce the reader to the concept of two kinds of mobilties:

First kind $\mu_{lattice}$ due to the scattering caused due to lattice thermal vibration and it is temperature dependent;

Second kind $\mu_{impurity}$ due to the scattering caused due to the dopents and/or crystalline defects;

At liquid Helium temperature that is at less than 4Kelvin there is no scattering due to thermal vibrations but scattering due to impurity and/or crystal defect persist. Therefore semiconductors never become superconductors. Even at 0Kelvin residual resistivity persists due to impurity and/or crystal defect.

So the effective mobility is given by the following reciprocal relationship:

$$1/\mu = 1/\mu_{lattice} + 1/\mu_{impurity} \quad \ldots\ldots\ldots\ldots\ldots \quad 1.113$$

The quantum-mechanical model of electron scattering , which has been presented in this chapter is valid only when drift velocity is much lower than the thermal velocity. When we reach high Electric Field region, the electric energy is directly transferred to the crystalline lattice and the drift velocity saturates at Scatter Limited Velocity as shown in Figure(1.72). The Scatter Limited Velocity is 10^7 cm/sec for Silicon and it can be derived from the following relation:

$$(1/2)(m_e^*)(v_{scatter\ limited})^2 = (1/2)(m_e^*)(v_{thermal})^2 = (3/2)kT \ldots\ldots\ldots\ldots\ldots \quad 1.114$$

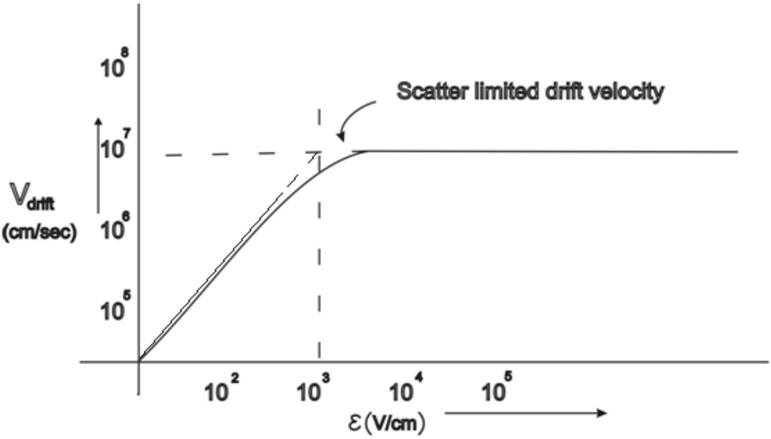


Figure 1.72. Vdrift versus Electric Field.

Figure 1.72. In high electric field region the drift velocity saturates at scatter limited velocity.

The Quantum Mechanical perspective tells us that electron is not impeded by the lattice centers. If the lattice centers are perfectly orderly at 0Kelvin then electron if imparted an impulsive energy will acquire a finite kinetic energy and with this KE it will continue to travel in a straight line through the crystalline lattice till infinity. What impedes the flow and causes the loss of KE is not lattice centers per se but the disorderliness of the lattice center network. This is the reason why Graphene is working out to be a wonder material with a very large drift mobility. Graphene is a sheet of orderly arrangement of hexagonal structure which is unperturbed or unbroken over large distances hence mean free path in Graphene is of the order of micrometers as compared to the mean free path in GaAs where it is of the order of a fraction of micrometer.

SSPD_Chapter 1_Part 13_ Constituent particles of Solid State of Matter_Their Birth and Evolution. SSPD_Chapter 1_Part 13 traces the birth and evolution of the Universe and how matter evolved since the beginning.

SSPD_Chapter 1_Part 13_ Constituent particles of Solid State of Matter_Their Birth and Evolution.

1.13. THE FUNDAMENTAL CONSTITUENT PARTICLES OF OUR UNIVERSE AND THEIR STATISTICS- BOSE-EINSTEIN STATISTICS AND FERMI-DIRAC STATISTICS.

From the beginning of proto-human and human society, “how ,why and where from” like questions have been arising in human minds.

About 8 million years before the present, plate tectonic movement led to the formation of the Great Rift Valley in Africa specifically in Tanzania. The West Side climate was unaffected and it remained as it was before - lush, green forests. The inhabitants on the West side grew as they were growing. On the West Side the Primate group of hominids (tailless monkeys) were developing into Chimpanzees and Apes. But the East Side underwent a drastic climatic change. Dry and arid condition transformed the lush, green forest covers into large savannah grasslands with clumps of trees few and far between. On the East Side , hominids developed into

Prognids. These arboreal creatures which lived in the trees were confronted with completely new challenges.

If we trace our genealogical path of evolution, old world monkey developed from squirrel monkey also known as insectivore. A phylo-genetic branch of old world monkey developed into a species of tail-less monkey.

On the West Side of the Great Rift Valley, these tail-less monkeys evolved into chimpanzees also known as Panidae.

On the East Side, these developed into Hominidae.

In palaeo-anthropolgy (the study of Proto Human Societies), the first form of hominidae is referred to as AUSTRALAPITHECUS (the southern apes). Hominidae is completely distinct from Apes and Chimpanzees.

The fruits and flowers became scarce and the males of this genre had to come down the trees in search of food particularly roots and underground tuber. They formed small bands of food gatherers and moved along Omo river in search of food. Once enough food was collected they carried them in their hands to their arboreal abodes, So by necessity they became bipeds whereas their female counterparts in the trees remained quadraped.

Thus a genre of tail-less hominids developed into erect , walking bipeds which were referred to as Austrlopethecus Afarensis.

These developed into Homo-habilis (handy man), Homo-erectus (erect man) and Homo-sapiens (wise man).

The roots and fruits gatherers became scavengers. The discovery and use of fire turned these scavengers into powerful hunters , tool makers and cave dwellers.

Probably until 30,000 years B.P. he was a nomadic hunter living in caves. He was able to make tools and weapons out of bone, wood and flint stone.

His knowledge from experience, known as empirical knowledge, was recorded in the inner parts of his cave as wall paintings. Many such wall paintings are being discovered in pre-historical cave settlements in different parts of the world. These wall paintings shed light on the life style and the material achievements of those cave dwellers. Proto-human and human societies during this period are referred to as Savage Society and subsequently as Barbarian Society. Human beings were nomadic because they were collectors, foragers and hunters.

At the end of the last Ice Age which was marked by Great Floods about 10,000 year BP a qualitative change occurred in the Human society. In course of hunting, human beings learnt husbandry, cattle rearing and in the process of growing fodders for their livestock they also learnt agriculture. This completely changed their lifestyle from that of a nomad to that of a settled village life. This marked the beginning of Civilization.

This also saw the emergence of State, Private Property and Monogamous Patriarchal Family. By corollary the Savage and Barbarian Society had self rule, community property and matriarchal gen families.

About 7000 years BP a kingdom under Manu Clan was thriving in the Gangetic Valley and Brahmputra Valley. Almost simultaneously a great City State was thriving in the river valleys of Sapt Sindhu. This we know as Mohenjaro and Harappan Civilization. This expanded along the legendary Sarswati River, which is completely lost today, upto Rann of Kutch.

Almost simultaneously Mesopatamian Civilization thrived in the river valley of Tigris and Eupharates (in modern Iraq).

In the river valley of Nile River , Egyptian Civilization thrived and prospered.

In the river valley of Yangtze and Yellow River(the river of sorrow), the Chinese Civilization developed.

The development of production and productivity in these River Valley Civilizations led to increased surplus value. The increased surplus value led to capital formation and exchange. Trade and Commerce centers developed.

Trade and Commerce Centers led to the development of city states. Development of such city states on Greek Peninsula led to the development of Greek Civilization.

Greek Philosophers were not satisfied with the answers to the question ‘HOW’. They wanted to know the theoretical basis of all the events in Nature. They wanted the answers to the questions ‘WHY’ and ‘WHEREFROM’.

The question raised in this Modern Age were all raised in the Ancient Age of the Greek Civilization.

In BC450, two Greek philosophers had proposed a hypothesis for the birth and evolution of Universe which was very near the modern theory of Cosmogny and Cosmology. But the stalwarts of those times, namely Plato and Aristotle outright rejected these hypothesis.

Greek Civilization offered the Ptolemic World View which regarded Earth as the center of our Solar System. This was called the Geocentric World View.

During the renaissance period , Copernicus and Galileo had to make extreme sacrifices and face excommunication for putting forward the heliocentric world view which puts Sun at the center of our Solar System.

In Rig Veda of Indian Civilization many hypothesis were put forward regarding Cosmogny and Cosmology.

In Bible a very definitive scenario is described regarding the theory of Creation which is known as GENESIS.

According to Genesis, 4000 years BP the Universe was created by God. In first six months the creation of all inanimates as well as animates was achieved. Since then every thing has remained immutable and unalterable.

In 19th Century and 20th Century as Sciences of rocks , fossils and celestial bodies developed, the Creationist Theory was questioned.

Through radio-active dating it was confirmed that rocks are much older than what is supposed by the Creationists.

Second jolt was given by Wegner’s plate tectonic theory which propounded that the outer crust of Earth is divided into 13 plates which are floating around on the semi-molten upper basaltic mantle. Therefore the distribution of the continents were not always as they are distributed now. The distribution is ever changing.

The final jolt was given to the creationists by Expanding Universe Theory discovered in 1920.

In the Mount Wilson Observatory, Los Angeles, California, two scientists Hubble and Humason discovered that all distant galaxies seem to be receding from us. Further the recession velocity was directly proportional to the distance of the galaxy and the constant of proportionality was called H (Hubble constant). This theory of Expanding Universe gave birth to the concept of The BIG BANG THEORY.

1.13.1. THE BIG BANG THEORY.

In 1940 George Gamow propounded the theory of Big Bang. According to this at a singularity a big bang occurred in the remote past. This event occurred 13.7 billion years before the present. From that point onward space started expanding and time started ticking. At the time of Big bang, there was infinite temperature and infinite density at the point of singularity. We do not know what the form of matter was at that time. As space expanded and time passed , temperature came down and density reduced.

Till 10^{-43}seconds after the instant of Big Bang, quantum gravitation theory was applicable. This state is known as Super Symmetry. At this point all the four forces merge as one force as shown in Fig. (1.73).

The four forces present in nature are: Gravitational force, Electromagnetic force, Strong force and Weak force.

In Table(1.27) , a quantative description of the four forces are given. The radius of the sphere of influence and its magnitude at one Fermi distance (10^{-15}m) is enumerated in the Chart. The radius of the nucleus is Seven Fermi Unit but was measured as 30 Fermi Unit by Rutherford Experiment.

Table 1.27 A comparative study of the four forces. [Table 13.1 from Chapter 13,'Elementary Particles', Concepts of Modern Physics, by Arthur Beiser, TMH, 6 th Edition, 2003]

Force	Radius of sphere of influence	Magnitude ofForce at UnityFermi distancein units of strong forceat	Particlesaffected	ParticlesExchanged	Role in Universe

		1 Fermi distance.			
Gravitational	Infinite	10^{-39}	all	Gravitons	Assembles Matter into planets , stars and galaxies
Elecromagnetic	Infinite	$10^{-2} \sim 1/137$	Charged particles	Photons	Determines structures of atoms, molecules, solids and liquids;Is important in factor in astronomical universe;
Weak	Less than 10^{-18}m	10^{-5}		Intermediate Vector Bosons andW^{+}, W^{-} & Z^{0}	Mediates quarks & leptons; helps determine compositions of atomic

Strong	Less than 1 Fermi dist	1			Mesons	nuclei; Holds quarks together to form nucleons;

Between material particles gravitational force act. Between charged particles electro-magnetic force come into picture. The acceleration of charged particles give rise to electromagnetic radiation. In radio-active decay weak forces act. In the nucleus the strong force come into play among the nuclear particles. As seen from the Table(1.27), presently there is a great asymmetry among the four forces. But at the time of Big Bang up to 10^{-43}seconds the four forces were of equal magnitude as shown in Figure(1.73). All four forces have gauge symmetry mediated by gauge particles. This gauge symmetry comes from theory of Quantum Electrodynamics also known as Relativistic Quantum Field Theory. James Clarke Maxwell established that electric force and magnetic force are aspects of the same electromagnetic force. Salem – Weinberg showed that electromagnetic force and weak force are synthesized in Grand Unified Field Theory (GUT). Weak forces have hidden symmetry.

According to Particle Physics, *hadrons (baryons and mesons) participate in strong interactions and experience strong force and leptons (electrons and* *neutrinos) interact weakly. The hadrons are composed of quarks. The triplet of quarks are baryons namelyNeutron, Proton, Lambda and Sigma and doublet of quarks are Mesons(pion and kaon).*

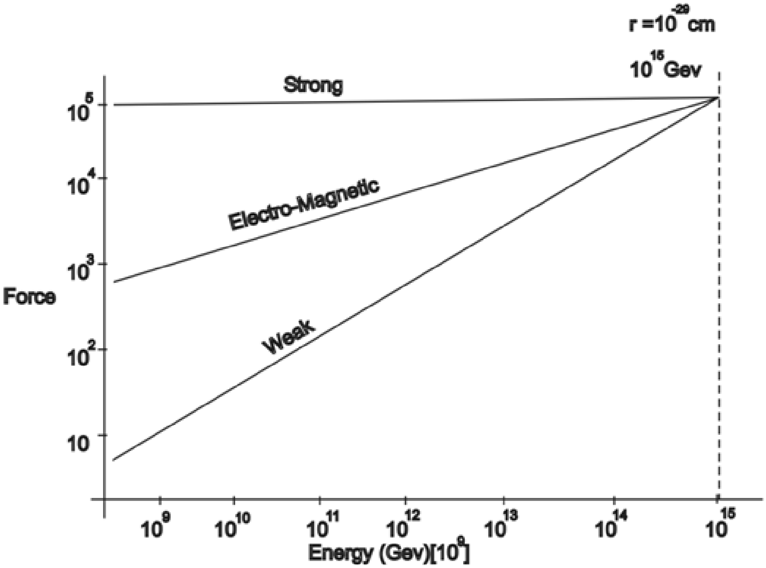


Fig.1.73 Comparative plot of Strong,Weak and Electromagnetic Forces at different ambient temperatures.

Fig.(1.73). Comparative plot of Strong, Weak and Electromagnetic Forces at different ambient temperatures.

At 10-43seconds after the Big Bang, the radius of the Universe was 10-33 cm and the density was 1093 gm/cm3. At this density matter was in the form of fundamental particles that is in the form of leptons and quarks. In Table(1.28), the charge and the rest mass are enumerated for the fundamental particles.

As seen from Table(1.28), quarks have fractional charge but in observable Universe we always find an integer number of electron charge. This clearly shows that quarks never find independent existence . *Pairs ofQuarks are always confined to the nuclear particles or messenger particles to give (±) unit electron charge or zero charge. Triplets of quarks create Baryons such as n and p with (±)1e charge or zero charge but always with ½ or 3/2 spin* . We have been able to achieve more than unity Tera eV. at Brookhaven National Laboratory and observed free quarks They exist as liquid with very little viscosity and not like gas particles.

Table 1.28. Comparaive study of Leptons and Quarks. Two groups and three generations of particles.

Lepton	symbol	Rest mass(MeV/c2)	Charge	Quarks	Symbol	Rest mass(MeV/c2)	Charge(q)
Neutrino	ν_e	?	0	Up	u	310	2/3
Electron	e	0.511	1	Down	d	310	1/3
Mu-neutrino	ν_μ	0	0	Charm	c	1,500	2/3
Muon	μ	106.6	1	Strange	s	505	1/3
Tau	ν_τ	164	0	Top	t	>22,500	2/3

neutrino							
Tau lepton	τ	1777	-1	Bottom	b	5,000	-1/3

* All particles have their anti-matter counter part.

Indirect proofs have been obtained for 3 families of particles. In 1992 , CERN, Geneva, and SLAC, Stanford, have conclusively proved that there are only 3 families of particles.

Theory says that if there are N generations of particles then Z_0 (intermediate vector boson mediating weak forces) have N modes of decay and will have N lifetimes.

Since only 3 decay modes and 3 lifetimes have been observed after investigating 4000 events it has been conclusively proved that there are only three generations of particles present in this Universe.

First generation consists of electron, neutrino and up- down quark. Second generation consists of mu electron or muon, mu neutrino and charm-strange quark. The third generation consists of tau lepton, tau neutrino and top-bottom quark.

The mass of electron-neutrino and mu-neutrino have been shown to be undetermined and zero respectively. Recently some new light has been shed on the rest mass and this is going to be crucial in determining the ultimate fate of the Universe.

On March 30, 2010, Large Hadron Collidor at CERN, Geneva, two oppositely rotating proton beams at 7 TeV were collided. Subsequently lead ion beam will be collided. This will help determine strong interaction and evolution of matter in early Universe. Precise measurement of CP violation and rare decays will be made. The hunt will begin for dark matter which is abundant in present day Universe. Some of the major puzzles of Modern Physics will be addressed like the origin of matter in hadron and the grand unification of forces. Standard Model will be rediscovered . This will be the necessary precursor to looking for New Physics.

Our galaxy known as Milky Way belongs to a cluster of galaxies known as Local Group. Belonging to the same Local Group is another galaxy known as Large Magellenic Cloud (LMC). In this galaxy on 23rd February 1987 there was a supernova explosion in a star named 1987A. This supernova explosion was discovered by Ian Shelton who was a technician but after this discovery was admitted into Graduate School.

A main sequence star lighter than 1.44 Solar Mass (ChandraShekhar Limit) after completing its life span explodes into Nova Explosion and expands into Red Giant and finally contracts into a White Dwarf.

A star heavier than 1.44 Solar Mass and lighter than 3 Solar Mass explodes into Super-Nova Explosion and the star contracts into a Neutron Star. The neutron star has a magnetic dipole . This magnetic dipole axis is inclined to the geographical polar axis and this leads to synchrotron radiation which makes the neutron star a pulsar.

A star heavier than 3 Solar Mass also explodes into Super-nova explosion but the star collapses into Black Hole.

After 1987A event scientists got an opportunity to study neutrino and mu-neutrino extensively and it has been decisively established that they have a finite mass .

This mass tells that our Universe is flat Universe and it has zero curvature. An open Universe will go on expanding up to eternity and ultimately all celestial objects will become cold and radiationless.

From Newtonian Mechanics we know that the escape velocity is given by the following equation:

V escape = √(2GM/R) ……………………………………. 1.114

Where G = Gravitational Constant;

M = mass of the body under consideration;

R = radius of the body under consideration.

Hubble and Humason had established that all the galaxies are racing away from one another and the velocity of recession is directly proportional to the distance of the galaxy from our Earth:

V recession = R.H.. 1.115

Where H = Hubble Constant and 1/H = Age of the Universe which has been accurately determined to be 13.7Gyears.

If the velocity of recession of the observable horizon (that is the velocity of recession at 13.7 G light years) is $V_{recession}$* then we have three possible scenarios for the final fate of this Universe:

V recession * = V escape : Flat Universe with 0 curvature;......................... 1.116

V recession * < V escape : Closed Universe with (+) ve curvature;.................. 1.117

V recession * > V escape : Open Universe with (-)ve curvature;....................... 1.118

A closed Universe with positive curvature is like a sphere which will ultimately stop receding and collapse on to itself in a final crunch and another big bang leading to the beginning of a new Universe.

A open Universe with a negative curvature is like a saddle which will continue to expand indefinitely. Ultimately everything becomes lifeless, radiationless , dark and cold. It will continue to expand with a whimper.

A flat universe has zero curvature. Here there is steady expansion and no accelerated expansion. This also will end in a cold death.

All these three scenarios have been shown in Figure 1.74.

If we take the flat universe scenario we arrive at the critical mass density $\Omega_{critical}$.

For Flat Unverse: $V_{recession}* = V_{escape}$

Therefore: RH= √(2GM/R)

Squaring both sides: $(RH)_2 = (2GM/R)$

Therefore: **Ωcritical = M/{(4π/3)R^3} = mass density of the Universe**

= 3H ^2 /(8πG) = 10 ^-29 gm/(cm) ^3 = 10 ^-26 Kg/(m) ^3 .... 1.119

where H = 75 Km/(sec.Mparsec);

$G = 6.67 \times 10^{-11}(m^3/(sec^2.Kg))$.

$\Omega_{actual}/\Omega_{critical} = \Omega_0 =$ normalized density of the Universe;

The critical density corresponds to 5 Hydrogen atoms per cubic meter.

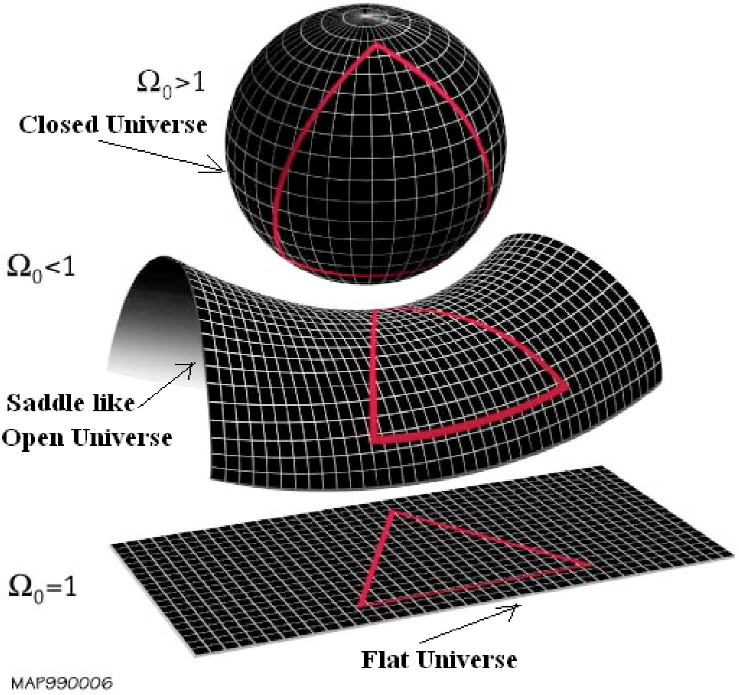


Figure 1.74. The three possible spatial geometries of the Universe

Fig. 1.74. The three cosmological models of Universe from the equations of General Theory of relativity.

In Table(1.28), we find that Top quark has the heaviest rest mass, greater than 22GeV/c^2. Therefore it was the last quark to be experimentally verified. In February 1994, it was finally confirmed with definiteness. So it is now a well established fact that there are only three generations of fundamental particles.

1.13.2. THE CREATION OF MATERIAL PARTICLES SUBSEQUENT TO THE BIG BANG.

If we consider the instant of Big Bang to be zero second then during the interval between 0 second and 10^{-43}second , we have no Physics to explain the events occurring in the given interval. Quantum Gravitation Physics will have to be developed to explain things.

Planck's time t = (Għ/c ˆ5) ˆ1/2 = 10 ˆ-43 seconds.............................. 1.120

At this instant ,

the radius of the Universe=(Għ/c ˆ3) ˆ1/2 = 10 ˆ-33 cm.......................... 1.121

kT gives temperature in eV. 0.86×10^{-4}eV corresponds to 1 Kelvin temperature increment;

Planck's Temperature:

T*(MeV) = 1/√t = 1/√(10 ˆ-43) = 3.16227×10 ˆ21 MeV~10 ˆ19 GeV...................... 1.123

T(Kelvin) = 10 ˆ10 /√t = 3.16227 × 10 ˆ31 K;**

Planck’s Density of the Universe = ρ planck = 10 ˆ93 gm/(cm) ˆ3 .... 1.124

Atomic Nucleus density = ρ atomic = 10 ˆ14 gm/(cm) ˆ3

From the difference of the two densities, we conclude that matter is not in normal condition at Planck’s Time. In fact it is being created from Photon every moment and instantaneously it is being destroyed:

Photon + Photon ↔ Particle + anti-Particle................................... 1.125

At Planck’s time, the temperature is 10 ˆ19 GeV

and the wavelength of the photon is: λ = [1.24/E(eV)] μm

= [1.24/(10 ˆ19 × 10 ˆ9 eV)] μm

λ = 1.24 × 10 ˆ-34 m 1.126

As we have already shown that at the super temperature of 10^{19}GeV, all the four forces are unified. This is called the super symmetry state.

With further expansion, temperature falls below

10^{19}GeV and super symmetry is broken. Here the Gravitation force decouples and Universe undergoes a phase transition from Quantum Gravitation to Grand Unified Phase. The Graviton Bosons which decouple are frozen and referred to as relic Gravitons. The Gravitons are boson particles which mediate the gravitational force.

We know from Classical Physics that if water undergoes phase transition from gas to liquid or from liquid to solid, it gives off latent heat without effecting the temperature. In a similar fashion, during the phase transition from Quantum Gravitation to Grand unified Phase, a large energy is released which changes the monotonic expansion to inflationary expansion.

Allen Guth gives an alternative conjecture for inflationary expansion:

In gauge field it is possible for vacuum to get excited. It possesses enormous pressure and energy. This pressure comes out to be negative. Normally mass , energy and pressure gravitate. In case of false vacuum, pressure causes three times as much energy and being negative it causes anti-gravity i.e. expansion.

Guth reasoned that if the Universe originated in an excited vacuum state, it need not have begun with a bang. It in fact started with a quantum whimper. Antigravity launched Universe on an inflationary phase-

exponentially growing rapid expansion with e-folding time of 10^-35second . In a dozen e-folding times, whimper became a bang and any magnetic monopoles if existing became so diluted in density that they became traceless.[New Physics- A Synthesis, by Paul Davies, taken from the book " The New Physics" edited by Paul Davies, Cambridge University Press, 1992]

Recent developments in Astronomy and Astrophysics have completely vindicated Allen Guth conjecture.

Wilkinson Microwave Anisotropy Probe (WMAP) gave a detailed picture of 3Kelvin Cosmic Microwave Background Radiation. This picture clearly indicates that Universe is only 4% ordinary matter made of electrons, protons and neutrons, 23% is dark matter made of as yet undetected exotic particles and 73% is dark energy or antigravity energy which Allen Guth had used in his conjecture. Ordinary matter and dark matter are responsible for gravitational attraction whereas dark energy is responsible for anti-gravity push which is responsible for accelerated expansion of the Universe as presently observed. WMAP data also establishes that the density of mass is just equal the critical density as a result the Universe is flat.

In the past decade Sloan Digital Sky Survey (SDSS) has been carried out in which a million galaxies have been mapped. The analysis of that data also arrives at the same conclusion that Universe is dominated by dark energy.

In July 2003, Scientists superimposed SDSS data on the microwave intensity map developed by WMAP. They conclusively proved that there is what is known as integrated Sachs-Wolfe effect. Clustering of galaxies cause dimple in space- time fabric. Anti-gravity outward push tends to stretch out the dimples in an otherwise flat Universe and in the process tends to crush the Cosmic Microwave Background Radiation at and near the dimples meaning by anti-gravity push shifts the microwave radiation towards shorter wavelength. This precisely was observed in the superimposed map. This is a definitive proof of dark energy dominance in our present Universe.[‘Breakthrough of the Year’, Scientist, 19th December , 2003, Vol. 302, pp.2038-2039]

From 10^{-43} second to 10^{-35}second, Universe experienced inflationary expansion. Eq.(1.123), which gives the temperature, is applicable only after the inflationary phase.

According to Eq.(1.123), 10^{-10}second after the big bang temperature reduces to 100GeV. Below this energy, the weak force freezes and this is known as GUT(Grand Unified Theory) Phase to Electro-Weak Phase transition . This Phase Transition is known as Salam-Wienberg Transition. As a result of this transition, weak forces decouple and Intermediate Vector Bosons, W_+, W_- and Z_0 freeze out as relic Vector Bosons.

Salam and Weinberg are two scientists who worked on electro-weak theory. Abdus Salam is of Pakistani origin who predominantly worked in West. Steven Weinberg is an American Scientist who wrote the popular book “ First Three Minutes”. These two Scientists were able to unify the weak force with electro-magnetic force and predict the existence of intermediate vector bosons which mediate the weak forces. For this discovery they were awarded Physics Nobel Prize in 1979 whereas the actual existence of intermediate vector bosons were proved in 1983. This is the reason why the transition from GUT Phase to Electro-Weak Phase is referred to as Salam-Weinberg Transition.

What are Vector Bosons?

W_+ & W_- are messenger particles mediating the β decay. These have rest mass $m_W = 85m_P = 80GeV$.

Z_0 mediate the strong force events. These have rest mass $m_Z = 97m_P = 91GeV$.

At some primitive level that is at energy levels much higher than 100GeV, both forces, that is weak and electro-magnetic forces, are aspects of a single interaction mediated by four massless bosons.

At 10^{-43} seconds after the Big-Bang the first spontaneous breaking of symmetry took place during Quantum-Gravitation to GUT phase transition. At that point symmetry breaking GRAVITATIONAL FORCE decoupled with relic gravitons freezing out.

At 10^{-10} seconds after the Big-Bang, the second spontaneous symmetry breaking takes place during GUT to Electro-Weak phase transition. At that point WEAK FORCE decouples and relic intermediate vector bosons freeze out.

After 300,000 yrs after the Big-Bang, ELECTROMAGNETIC FORCE decouples and relic photon or relic radiation freeze out which due to Hubble Expansion have today cooled down to and observed as 3 Kelvin Cosmic Microwave Background Radiation (CMBR).

At the second spontaneous symmetry breaking, out of the four massless bosons three acquire masses and become W^+, W^- and Z^0 Bosons.

Bosons which are massless have infinite range of influence such as photons. But bosons, the messenger particles, which have masses have their range of action drastically reduced. The examples are Intermediate Vector Bosons which have rest masses of the order of 90GeV with a range of action of the order of 10^{-18}m and K^+ mesons, with rest mass 815MeV, have a range of action of 1 Fermi i.e.10^{-15}m.

The title of this section is the fundamental constituent particles and their statistics. As we have already seen in SSPD_Chapter 1_Part 9_conclusion there are two statistics- Bose-Einstein Statistics and Fermi-Dirac Statistics.

The two statistics give two distinct distribution functions. The distribution function is the probability of finding the particles at different energy levels. The particles which obey Bose-Einstein Statistics are called Bosons(indistinguishable particles and have symmetric wave function) whereas the particles which obey Fermi-Dirac Statistics are called Fermions(distinguishable particles and have anti-symmetric wave functions).

Electrons, Protons and Neutrons are Fermions. Fermions always posses spin angular momentum equal to the integral multiple of (1/2)ħ. Fermions are real material particles and obey Pauli-Exclusion Principle. An elemental momentum phase space contains only two opposite spin Fermions. Because of Pauli-Exclusion Principle, if more than two fermions are squeezed in an elemental volume there will be strong repulsive force. Fermions behave claustrophobically that is two Fermions cannot occupy identical Quantum States in the same

Cartesian Elemental Space. Electrons, protons and neutrons behave claustrophobically.

Contrary to these the carrier particles of the four forces are Bosons. Bosons are gregarious. They tend to gather in identical states if given the opportunity. An atom or a molecule with even number of e’s, p’s and n’s are bosonic in nature.

The carrier particles of gravitational force are gravitons .

The carrier particles of weak forces are intermediate vector bosons.

The carrier particles of electromagnetic force are photons.

The carrier particles of strong forces are Mesons (k Mesons and π Mesons).

These particles obey Bose-Einstein Statistics and always posses spin angular momentum equal to 0ħ, 1ħ, 2ħ ……. A large number of Bosons can be accommodated in a very small volume and these particles will experience attraction. In this sense these are virtual particles.

Now we return to the expanding Universe.

As the energy level comes below 100GeV, the weak force freezes and intermediate vector bosons decouple from the hot soup and relic vector bosons are left over to this day.

Now the hot soup is an admixture of leptons , quarks and photons and only strong force and electromagnetic remain unified. Gravitational force was first to decouple leaving behind relic gravitons. Weak force was next to decouple leaving behind relic vector bosons.

At 10^{-6} sec the energy fell to 1GeV. This energy is equal to the energy equivalent of proton rest mass.($E=mc^2$: this is the celebrated equation given by Einstein in his Special Theory of Relativity on which the fission bomb, more popularly known as atom bomb, and fusion bomb commonly known as hydrogen bomb are based)

That is $E= 939.57\text{MeV} = m_p \times c^2 /q$;

This implies that before 10^{-6} sec. protons and neutrons are being fissioned into their constituent quarks:

Neutron split into 2 Down quarks + 1 Up quark;

Proton split into 2 Up quarks + 1 Down quark………………… 1.127

From Table(1.28) we find that 2 Down quarks + 1 Up quark amount to 0 coulomb of charge and $(930+\Delta)$MeV of equivalent mass. Therefore

Neutron has 0 charge and $(930+\Delta)$ MeV of equivalent mass.

Similarly 2 Up quarks + 1 Down quark amount to 1q amount of charge and 930MeV of equivalent mass. Therefore Proton has $+q$ coulomb of charge and 930MeV of equivalent mass.

This is precisely why when the temperature of the Universe is greater than 1GeV then Proton and Neutron do not have an independent existence and they exist as a soup of gluons and quarks.

Quarks interact through gluons. Gauge Theory of inter-quark force is Quantum Chromo-Dynamics(QCD). Gluons play the same role with respect to inter-quark force as photons play with respect to electromagnetic force. Quarks combine to form Mesons which are the messenger particles of Strong force. These Mesons have charge which are called color hence the word chromo in QCD.

The unification of Strong, Weak and Electro-Magnetic force above 100GeV is the Grand Unified Theory(GUT). GUT is tested through the existence of Magnetic Monopoles. In classical electrodynamics the existence of magnetic monopoles is precluded. Therefore the DIVERGENCE of B is always zero in four Maxwell Equations.

But DIVERGENCE of D is zero only when Cartesian Space under consideration includes no space charge or equal amount of negative and positive charges otherwise DIVERGENCE $D = \rho$ (net charge density). DIV $D = 0$ is known as Laplace's Equation and DIV $D = \rho$ is known as Poisson's Equation.

GUT is also being tested through the decay of protons.

Unification of Gravitational force along with GUT i.e. unification of Gravitational Force, Strong Force, Weak Force and Electro-Magnetic Force is Theory of All (TOA) or the SUPER-SYMMETRIC STRING THEORY.

Within the nucleus, which has a dimension of 7 Fermi($\times 10^{-15}$m) units, protons in spite of repulsion coexist along with neutrons because of strong force which is several orders of magnitude greater than electromagnetic forces. The carrier particles of this force within the nucleus is Mesons (bosonic particles). Therefore strong forces are also referred to as Mesonic forces. Above 1 GeV, just like protons and neutrons, mesons also loose their integrity and split into the following quarks:

K + Mesons ↔ Up quark + Strange anti-quark;

Π + **Mesons ↔ Up quark + Down anti-quark; ………………… 1.128**

In Table(1.28), no mention is made of anti particles. In fact all material particles have anti-particles.

Electron has Positron as anti-particle. Anti-particles have the same mass and spin but opposite charge. Positron and Electron have identical masses but opposite charge.

Similarly Proton has anti-Proton. The two have identical masses but opposite charge.

All three generations of quarks, leptons and neutrinos have their anti-particle counterpart. When two material anti-particles combine they get annihilated to give equivalent amount of energy by mass equivalence law given by special theory of relativity:

$e + \text{anti } e \leftrightarrow 2\gamma$ (soft gamma photons)

$p + \text{anti } p \leftrightarrow 2\gamma$ (hard gamma photons)

	Up quark	Strange antiquark	K_+ Meson
Charge	$(2/3)\ e_0$	$(1/3)\ e_0$	e_0
Equivalent mass	310MeV	-505MeV	-195MeV
Spin	$(1/2)\hbar$	$(-1/2)\ \hbar$	$0\hbar$

……………….1.129

	Up quark	Down anti-quark	Π+
Charge	(2/3) e_0	(1/3) e_0	+ e_0
Mass	310MeV	310MeV	0 eV
Spin	(1/2) ħ	-(1/2) ħ	0 ħ

……………………….1.130

From Eq.(1.129) and Eq.(1.130) we see that both K + and Π+ have 0ħ spin angular momentum hence they are Bosons. But K+ has mass equivalence of -195MeV and Π+ has mass equivalent of 0eV. Therefore in a nucleus where equal mass particles are present there Π+ mediate as carrier particles but where unequal mass particles are present there K+ mediate as carrier particles.

When particle and antiparticle meet they annihilate and mass equivalent amount of energy is released as photon of appropriate wavelength.

Consider electron and positron: $m_e c^2/q = 0.5$MeV;

Therefore e + e(anti) = 1MeV↔ 1.24/(1×10 ˆ6 eV) = 1.24×10 ˆ-6 μm = 1.24pm(pico meter)

Pico to Femto meter corresponds to gamma rays(γ rays).

Pico is soft gamma rays whereas Femto is hard gamma rays.

Consider proton and anti-proton: $m_p c^2 =$ 0.9396GeV;

Therefore p + p(anti) = (2 × 0.9396)GeV ↔ 1.24/(2 × 0.9396 × 10^9 eV) = 0.67 × 10^{-15} m = 0.66fm;

Thus we conclude that electron and its antiparticle positron collide to give soft γ photons and proton and antiproton annihilate to give hard γ photons.

Amongst the nuclear particles proton and neutron are the lightest(mass equivalent of 1GeV). Therefore at t = 10^{-6} sec the temperature is 10^{13}K which is equivalent to 1GeV. Therefore with expansion of the space-time fabric, temperature further falls. The hot soup of electrons-quarks-photons undergoes a phase transition to electron-nucleon soup. This is known as quark-nucleon phase transition. Below 1 GeV strong nuclear force freezes out, relic quarks get decoupled and we are left with a hot soup of hydrogen-helium nucleons, electrons and photons equivalent to 10^{13}K. This is known as the fourth state of matter namely plasma. In this hot plasma the photons are energetic enough to prevent Hydrogen and Helium from staying in deionized state or in neutral state . From t = 10^{-6} second (temperature 10^{13}K) to t = 1 second (temperature 10^{10}K) = 1000 times Sun’s core temperature , hydrogen-helium nucleons are continuously inter- transmuting between protons and neutrons i.e.

p ↔ n ……………………………………. 1.131

From t = 10^{-4} second (temperature 10^{12} K) to t = 1second (temperature = 10^{10}K) the inter-transmutation is leading to a steady state value of p (proton particles) : n (neutron particles) = 7:1.

After t = 1second i.e. below 10^{10}K, quarks condense to a steady value of p : n = 7:1.

The steady state value is given by:

$$\frac{[n]}{[p]} = Exp\left[-\frac{(\Delta m)c^{2}}{kT}\right] = \frac{15n}{100p} \quad at\ 10^{10}K \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 1.32$$

Where

$$\Delta m = m_n - m_p$$

After t = 100 seconds (1.7 minutes) temperature comes down to 10^{9}K and around 3 minutes after the Big Bang the elemental form of matter has the following composition: 25% Helium , 75% Hydrogen and traces of Deuterium and Lithium.

In the time interval of 1second to 180 seconds after the Big Bang:

p : n = 7: 1 or p : n = 14: 2 combine to give deuterium and hydrogen nuclei. Deuterium($_2D_1$ = 1 p + 1 n) is an isotope of hydrogen ($_1H_1$).

p + n → 2 D 1

2 D 1 + 2 D 1 → 4 He 2 ………………………………………………… 1.133

At the Deuterium stage there are 12 (${}_1\mathrm{H}_1$) hydrogen nuclei and 1(${}_4\mathrm{He}_2$) helium nucleus.

(p : n = 14:2) → (12 × ${}_1\mathrm{H}_1$) + (1 × ${}_4\mathrm{He}_2$)

Therefore weight ratio is :

H: He = 12 Atomic Weight : 4 Atomic Weight=75% : 25% ………………………………………………… 1.134

Therefore 3 minutes after the Big Bang, the primordial matter in Universe is in elemental form and most primitive form. The evolved form of matter come about by nucleosynthesis in the interior of stars up to Iron. Heavier elements are synthesized during Super nova explosion when a star beyond 1.44M 0 collapses into a neutron star or in a black hole.

(1/10,000) of the primordial matter is residual Deuterium. (1/100,000) of the primordial matter is He-3. A still smaller fraction (1/10×10 9) fused to Lithium.[Reference "The Evolution of the Universe" by Peebles, Schramn, Turner & Kron, Scientific American, October 1994].

η = baryonic number/photon number is the fraction which determines the percentage composition of primordial matter in terms of H, He, D & Li.

Presently there are 411 photons/cm^3 and 0.4 baryons/m^3. This gives $\eta = 1/10^9$. [Reference “Primordial Deuterium and the Big Bang” by Craig J. Hogan, Scientific American , December 1996].

1.13.3. 300,000 YEARS AFTER THE BIG BANG

In band theory chapter we have seen that ionization energy of hydrogen atom is 13.6eV.

Therefore as Universe cools from 10^{13} K to 1.6×10^5 K, matter remains in plasma state.

As temperature of the Universe falls below 10^{13} K, quarks undergo confinement. They either confine as Hadrons which are Fermions or as Gluons which are Bosons. Protons and Neutrons belong to hadrons group of nuclear particles and Mesons belong to gluon group of nuclear particles.

At 10^{-4} second after the Big-Bang, temperature falls to 10^{12} K and energy equivalent is 100MeV. At this time and temperature, atomic nuclii come into existence. The matter density of the Universe is 10^{14}gm/(cm)^3 which is the density of atomic nucleus also.

At the quark-nucleon phase transition ,10^{-6}second after the Big Bang, the quarks which got decoupled should be there as relic or residual quarks and their mass abundance should be as high as gold. But till date no relic quark has been detected.

From 10^{-6}second to 10^{13}second (300,000 yrs after the Big Bang), temperature fell from 10^{13}K to 4000K.

4000K per se corresponds to photon energy of 0.3eV. But there are sufficient high energy photons in the tail of Black Body Radiation at 4000K which keep the hydrogen atoms ionized. Hence above 4000K and upto 4000K all hydrogen atoms arte ionized plasma bathed in black body radiation flux.

Below 4000K, the energy is insufficient to cause the ionization of Hydrogen and Helium. Hence the hot plasma deionizes to form neutral H_2 77% by weight and neutral He 23% by weight.

At quark to nucleon phase transition, the nuclii of H_2 and He namely proton and neutron had been formed in the ratio 75% and 25%. After deionization, neutral H_2 and He form in the ratio 77% and 23%.

When neutral atoms were formed the interaction between neutral atoms and photon stopped. Radiation got decoupled. This decoupled radiation today 13.7 Gy after is seen as the Blackbody Radiation at 2.77K. This 2.77K black body radiation appears as Cosmic Microwave Background Radiation.

In 1960's AT & T, Bell Lab used a radiometer in a telescope that would track the early communication satellites namely Echo-1 and Telster. This radiometer was being tested by Arno A. Penzias and Robert W. Wilson. In course of this testing they detected a background noise in microwave spectrum. It corresponded to 2.776K Cosmic Microwave Background Radiation(CMBR). They were led to this idea by the news that Robert H. Dick had suggested in 1940 that a radiometer would detect this CMBR.

The discovery firmly established the Big Bang Theory. In 1978 the Nobel Prize in Physics was awarded to Arno A Penzias (American), Robert Woodrow Wilson (American) and P.L. Kepiteza (Russian) for the discovery of Cosmic Microwave Background Radiation.

For 300,000 years this relic radiation was coupled with matter through Compton Scattering and temperature of Black Body Radiation and temperature of Matter are equal though they have different adiabatic indices. Because of Compton Scattering, radiation should have the same density distribution as the matter density distribution. After 300,000 years when temperature fell below 4000K, the matter and radiation decoupled because matter had recombined and neutralized. Since then the two cooled independently. $T(\text{matter})\ \alpha\ 1/R_2$ and $T(\text{Radiation})\ \alpha\ 1/R$ but Black Body Radiation continues to carry the imprint of Mass Distribution at the time of decoupling.

Since then Matter has evolved in a very differentiated and non-homogeneous system of super-clusters, clusters, galaxies and solar systems hence at 300,000 year after the Big Bang, matter density distribution should have contained ripples of correct size which could eventually evolve in today’s variegated structure. Hence by studing the relic radiation and its ripples we can ascertain if indeed such a coupling existed 300,000 years ago.

Following are the stages through which the matter density ripples must have evolved: matter must have first segregated into super clusters, super clusters evolved into clusters of galaxies, clusters further sub-divided into galaxies. The turbulence within the galaxies gave birth to the stars and some stars had their solar systems too.

If matter density distribution would have had density ripples then radiation density distribution should have had similar ripples of the correct size and these ripples should be observable today also.

In 1990, a satellite called COBE (Cosmic Background Explorer) measured the spectrum and temperature variations in the radiation pattern. The spectrum was exactly as expected for 2.7K Black Body radiator. The temperature variation was 1 in 100,000. Till then CMB radiation had appeared as perfectly isotropic and isotropic radiation pattern cannot explain the galaxies, clusters and super-clusters formation. Therefore the achievement of COBE marked the correctness of the Big Bang Theory.

How did the uniform, featureless Big Bang explosion evolve into the galaxies and stars we see today?

13.7Gya Big Bang occurred. Explosion is uniform and featureless. During the inflationary phase of expansion matter is still uniformly distributed.

300,000 years after the Big Bang, COBE satellite show huge wispy cloud of matter, 500 million light years in length and larger, beginning to condense - this is verified by the discovery of ripples of matter in deep space.

The first galaxies must have formed at about 200 million years after the Big Bang.

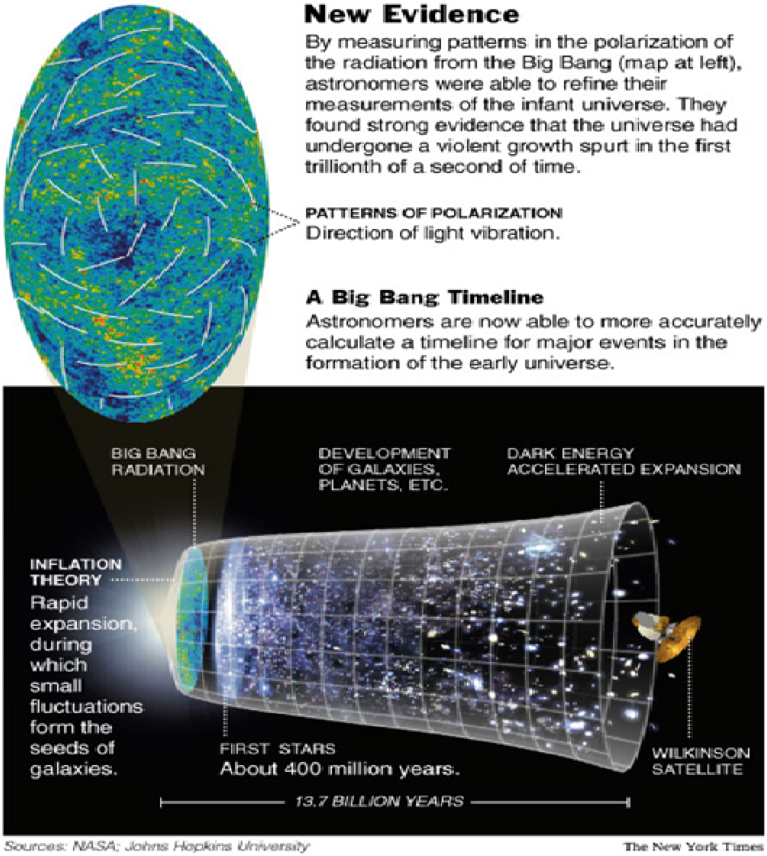
Today 13.7 billion years later, stars and galaxies have evolved out of the ripple of matter.

In 1993, right on the cue, came the news from NASA that ripples in the Cosmic Background Radiation of exactly the right size had been detected

–Dr. George Snout, the leader of the COBE team, announced at a meeting of the American Physical Society in Washington, what he described as evidence for the birth of Universe.

To refine the CMB radiation pattern studies, in 2001 Wilkinson Microwave Anisotropy Probe was launched (WMAP). This is in a circular orbit around Sun with semi-major axis $a = \; + \; 1.5 \times 10^6$ km.

In February 2003, the image of the infant cosmos only 400,000yrs old was received. The results from WMAP reveal that the CMB temperature variations follow a distinctive pattern predicted by cosmological theory: the hot and cold spots fall in characteristic sizes .The image of infant cosmos is shown in Figure 1.75.



New Evidence

By measuring patterns in the polarization of the radiation from the Big Bang (map at left), astronomers were able to refine their measurements of the infant universe. They found strong evidence that the universe had undergone a violent growth spurt in the first trillionth of a second of time.

A Big Bang Timeline

Astronomers are now able to more accurately calculate a timeline for major events in the formation of the early universe.

PATTERNS OF POLARIZATION
Direction of light vibration.
BIG BANG RADIATION
DEVELOPMENT OF GALAXIES, PLANETS, ETC.
DARK ENERGY ACCELERATED EXPANSION
INFLATION THEORY
Rapid expansion, during which small fluctuations form the seeds of galaxies.
FIRST STARS
About 400 million years.
WILKINSON SATELLITE
13.7 BILLION YEARS

Sources: NASA; Johns Hopkins University

The New York Times

Figure 1.75. The image of the infant COSMOS as taken by Wilkinson Microwave Anisotropy Probe and the Big Bang time line.

Hot spots (colored red) in the background radiation are the images of compressed, dense plasma in the cooling universe, and the cold spots (colored blue) are the signatures of rarefied region of gases.

WMAP give the following picture of the Universe:

1. 4% ordinary matter;
2. 23% exotic matter- the dark matter made up of an-as-yet-undetected-particle;
3. the remaining 73% is dark energy which is causing repulsion and hence accelerating the expansion rate of the Universe.

The age of the Universe is exactly pinned down at 13.7Gy ± 100,000yrs.

H, Hubble Constant, is exactly 71km per sec per megaparsec.

Universe’s shape is exactly flat, neither closed nor open. This means the density mass density is exactly at the critical mass density.

After 10^{-6}second the temperature falls below 10^{13}K, equivalent to 1 GeV, the quarks get confined into nuclear particles. This Quark-Nucleon Phase Transition. Two kinds of nuclear particles are created: hadrons which are Fermions and gluons which are Bosons.

Protons and neutrons are hadrons. Mesons are gluons.

After 10^{-4}second, temperature fell below 10^{12}K, equivalent to 100MeV. This provided the condition for the stable formation of nuclii of Hydrogen, Helium and Lithium. At this point the mass density of the Universe was 10^{14}gm/cc and this is the nuclear density also.

At the time of Quark-Nucleon Phase Transition, some quarks got decoupled and must exist as relic quarks in a given mass abundance. This mass abundance is the same as that of gold but till date no free quark has been detected.

From 10^{-6} second to 10^{13} second, temperature falls from 10^{13}K (equivalent to 1GeV) to 4000K(0.3eV). At this energy level the photons are not energetic enough to maintain the ionized plasma. The primitive form of ionized H and He and traces of ionized Lithium transform into neutral H_2 , He and traces of neutral Li. Once the matter neutralizes, energy present in form of electro-magnetic energy gets decoupled and starts cooling with continued expansion of the Universe and today it fills up the Universe as all pervasive 3Kelvin CMB (cosmic microwave background radiation).

We have already described that detection of CMB by WMAP has helped reveal the heterogeneity which must have been present in the matter density distribution which eventually led to the formation of Super-Clusters, Clusters, Galaxies and Solar Systems.

Our Solar System is 4.56 Gy old. It belongs to Milky Way Galaxy which is 10Gy old and is third generation galaxy. This Milky Way belongs to a cluster of galaxies known as Local Group. Local Group in turn belongs Virgo Cluster which in turn belongs to Virgo Super Cluster.

SSPD_Chapter1_Part 13_Conclusions of Big Bang. SSPD_Chapter 1_Part 13_ Conclusion gives a persepective view of Big Bang Theory. As the Universe expands and temperature cools the matter-energy stabilizes to a constant baryonic number. Conserving this baryonic number, all subsequent evolution follow-formation of super clusters, clusters, galaxies, solar-systems.

SSPD_Chapter 1_Part13_Conclusions.

1.13.3. THE CONCLUSIONS OF BIG BANG THEORY.

Table 1.29. The elementary particles of the Universe.

Carrier particles	Three generation of quarks and anti quarks	
Gravitons-carrier of gravitational forceRest mass- 0 GeV/c^2Jspin - 2ħClassified as bosons0 electric	up	Down

chargeThreshold energy- 3×10^{18}GeVThreshold temperature-3×10^{31}KAbove the threshold energy, super symmetry is present. All four forces are unified as a single force. Below the threshold energy, super symmetry breaks. Gravitational force decouples from the remaining three forces.

Photons- carrier of electromagnetic forcesRest mass- 0 GeV/c^2Jspin - 1ħClassified as bosons0 electric chargeThreshold energy- 0.3 GeVThreshold temperature-4000K. Above 4000K,	strange	Charm

radiation dominates. Below 4000K, matter dominates.

Intermediate vector bosons (W^+, W^- and Z^0) - carrier of weak forcesRest mass- 81 GeV/c^2, 81 GeV/c^2, 93 GeV/c^2 respectivelyJspin - 1ħClassified as bosonselectric charge positive, negative and zeroThreshold energy- 100GeVThreshold temperature-10^{15}KAt 10^{15}K (t = 10^{-10} second after the Big Bang), Salem-Weinberg Phase transition occurs. Above this temperature, strong force, weak and

top	bottom

electromagnetic forces are unified. Up to this temperature, Universe is in Grand Unification Phase also known as GUT. Below this temperature weak force decouples as an independent force and relic intermediate vector bosons freeze out. But strong force and electromagnetic force remain unified.Below 10^{15}K we are left with a quark, lepton and photon hot soup.

Hadrons(constituted of quarks)	
Baryons(triplet of quarks)Real particles,Jspin $= 1/2\hbar$, Constituent of	Gluons(doublet of quarks)Virtual Particles,Jspin $= 0\hbar$ Carri of strong

	nucleus, neutrons, protons Ξ^0, Ξ^-, Λ^0, Ω^-, Σ^-, Σ^+, Σ^0	Fermionic	forces, K mesons (K^0, K^-, K^+ and their antiparticles) and π mesons(π^0, π^-, π^+ and their antiparticles)	Bosonic.
Below 10^{10}K, all three generation of Leptons stabilize.	Leptons	Fermions	Spin=1/2ħ	Neutrinos
	electron			Electron neutrino
	Mu lepton			Mu neutrino
	Tau-lepton			Tau neutrino.

On 4th July 2012,the announcement was made regarding the definite bump at 125GeV which is consistent with Higg's Boson.The Higgs boson has been making headlines, but it is the corresponding Higgs field that is the true prize. Such fields are a fundamental part of modern physics, permeating space and interacting with the particles that move through them — as a magnetic field causes iron filings to line up.

In comparison with electromagnetic and other fields, however, the Higgs field is “slightly odd”, admits Tom Kibble, one of six theoretical physicists who first invoked the mechanism in 1964. The field is both ubiquitous and directionless, much like the constant air temperature in every part of a still, silent cave. Particles that interact with the Higgs field gain mass — the more they interact with it, the heavier they become.

Kibble, Peter Higgs and others proposed the existence of the field to solve one of the toughest physics problems of the day. In the early 1960s, theorists knew that two(unifiedly they were electroweak force) of the four fundamental forces(Gravitational, Strong, Electromagnetic and Weak force) that govern the behaviour of particles were nearly identical mathematically. The main difference between them was that the particles associated with one force(weak force) had mass, whereas those associated with the other(electromagnetic) did not.

The Higgs field explained the split. In the very early Universe, the theory goes, the Higgs field was zero, and the two forces were as one(unified electroweak force). But shortly after the Big Bang, the field assumed a non-zero value and the forces split apart(electromagnetic force and weak forces). One, which became electromagnetism, is mediated by massless particles of light known as photons, which ignore the Higgs field. The other force became the weak nuclear force, which causes certain kinds of radioactive decay, and works through heavy particles called intermediate vector bosons ,W and Z bosons. These interact with the Higgs field and gain mass. Ordinary matter derives most of its mass from subsequent interactions between particles such as quarks, found in the nuclei of atoms.

The Higgs boson itself can be thought of as an excited ripple in the Higgs field. Studying the boson’s properties at the Large Hadron Collider will tell theorists whether the field behaves as they expect.

The Higgs boson is an expression of the Higgs field — the mechanism ultimately responsible for the mass of known particles. The boson and field are needed in calculations to unify the electromagnetic and weak nuclear forces into a single ‘electroweak’ force, which in turn predicts the properties of other particles

The key quantity is the particle’s spin — a defining quantum-mechanical property. According to theory, the Higgs boson’s spin must be zero. Otherwise the masses of fundamental particles could change according to their orientations in space. “Spin zero is key,” he says. If the particle turns out to have a non-zero spin, it would be a shocking discovery — and would mean that the particle would be something other than the Higgs.

The latest LHC results reveal enough about how the particle decays to indicate that it has a spin of either zero or two. Further studies of the direction in which the particle’s decay products fly out of the

LHC will pin down the spin,

For the most part, the new particle matches the standard model's predictions of how it should decay into other particles. But there are some intriguing hints that this Higgs may not be entirely standard in character. For one thing, it seems to decay into pairs of photons about twice as often as the standard model predicts — and it decays less often than expected into particles known as taus and W bosons. Such discrepancies are far from being statistically significant right now, but if the evidence grows as more data are gathered, they could point to physics beyond the standard model.

For example, it may be that the detected boson is actually a composite, made up of smaller particles, or that it is the first of a new class of bosons — a 'Higgs family'. "I would be delighted if this new state is a Higgs boson, but perhaps not the standard-model Higgs boson," said Fabiola Gianotti, spokeswoman for ATLAS, at a press conference following the presentation of the results. Incandela thinks that by the end of the year, the experiments should have strong indications of whether the Higgs completely matches the standard-model predictions.

As the collider continues to smash protons beneath the hills of Switzerland and France, physicists hope that hints of the answers will emerge from the terabytes of data.

Till date no direct detection of gravitons has taken place. Scientist Weber of University of Maryland did detect gravitons in 1961 but it could not be repeated. The study of Cosmic Microwave Background Radiation(CMBR) will take us to a time 300,000 years after the Big Bang but not any earlier. In order to fully understand the birth and evolution of the Universe we need to go to the first split second after the Big Bang. In the first split second phase transition from Quantum Gravitation Phase to GUT Phase released enormous energy which fuelled the rapid inflation. The patch of Space that we now call the observable Universe started at subatomic size blew up to a few centimeter across in a fraction of a second. When large masses of matter are strongly accelerated then gravitational waves are radiated just as accelerated charge radiates electromagnetic waves. These gravitational waves are traveling warps in space time fabric. These carry signals from the dawn of time and have remained unaffected by intervening events. Just as matter density fluctuations, which ultimately blew up into galaxy seeds, have left their imprints on CMBR in form of hot spots and cold spots these gravitational waves have also squeezed and stretched the primeval fireball. By studying the polarization patterns in CMBR the imprints of gravitational waves can be separated from those of giant waves and from those of the matter density fluctuations. Determination of gravitational wave spectrum will help nail down the inflationary theory applicable to the birth of Universe or it may give a completely different theory of the birth of the Universe.[" The First Split Second", by Govert Schilling, New Scientist, 31st March 2001

The crux of Big Bang Theory has been enumerated in Table(1.29).

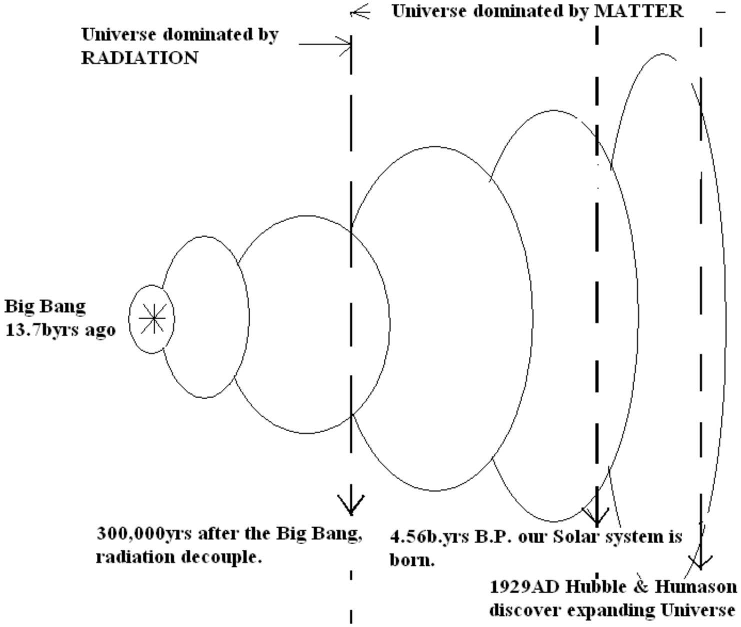

Universe dominated by MATTER
Universe dominated by RADIATION
Big Bang
13.7byrs ago
300,000yrs after the Big Bang, radiation decouple.
4.56b.yrs B.P. our Solar system is born.
1929AD Hubble & Humason discover expanding Universe

Figure 1.76.a. Different Phases of Expanding Universe after Big Bang.

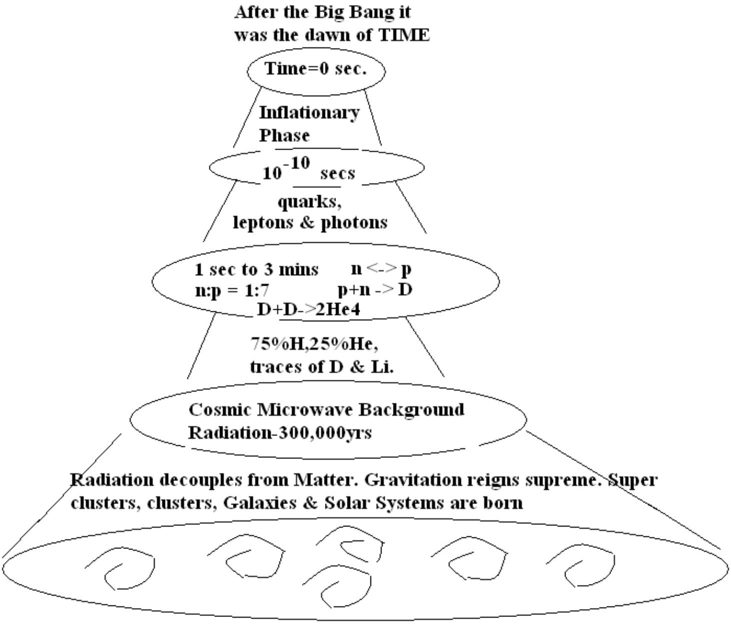

After the Big Bang it was the dawn of TIME
Time=0 sec.
Inflationary Phase
10^{-10} secs
quarks, leptons & photons
1 sec to 3 mins
n:p = 1:7
n <-> p
p+n -> D
D+D->2He4
75%H,25%He, traces of D & Li.
Cosmic Microwave Background Radiation-300,000yrs
Radiation decouples from Matter. Gravitation reigns supreme. Super clusters, clusters, Galaxies & Solar Systems are born

Figure 1.76.b. How the Universe Evolved.

[“First Split Second”, Govert Schilling, New Scientist, 31 st March 2001]

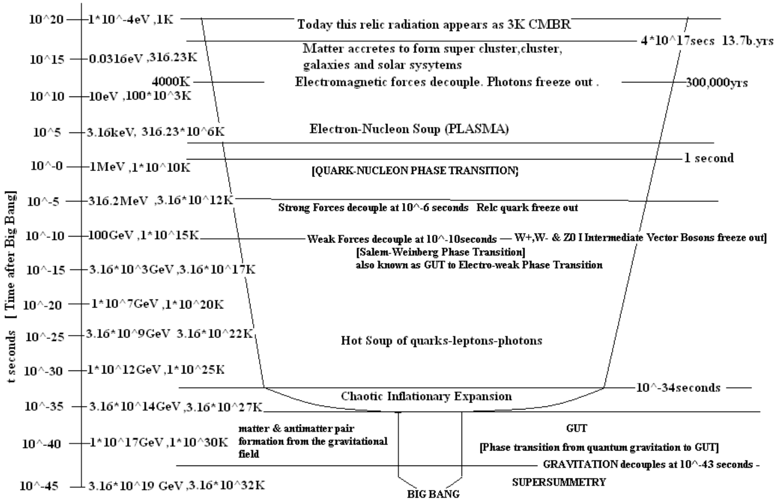

10^20
1*10^-4eV ,1K
Today this relic radiation appears as 3K CMBR
4*10^17secs 13.7b.yrs
10^15
0.0316eV ,316.23K
Matter accretes to form super cluster,cluster, galaxies and solar sysytems
4000K
Electromagnetic forces decouple. Photons freeze out .
300,000yrs
10^10
10eV ,100*10^3K
10^5
3.16keV, 316.23*10^6K
Electron-Nucleon Soup (PLASMA)
10^-0
1MeV ,1*10^10K
1 second
[QUARK-NUCLEON PHASE TRANSITION]
10^-5
316.2MeV ,3.16*10^12K
Strong Forces decouple at 10^-6 seconds Relc quark freeze out
10^-10
100GeV ,1*10^15K
Weak Forces decouple at 10^-10seconds — W+,W- & Z0 I Intermediate Vector Bosons freeze out]
[Salem-Weinberg Phase Transition]
also known as GUT to Electro-weak Phase Transition
10^-15
3.16*10^3GeV ,3.16*10^17K
t seconds [Time after Big Bang]
10^-20
1*10^7GeV ,1*10^20K
10^-25
3.16*10^9GeV 3.16*10^22K
Hot Soup of quarks-leptons-photons
10^-30
1*10^12GeV ,1*10^25K
10^-34seconds
Chaotic Inflationary Expansion
10^-35
3.16*10^14GeV ,3.16*10^27K
matter & antimatter pair formation from the gravitational field
GUT
10^-40
1*10^17GeV ,1*10^30K
[Phase transition from quantum gravitation to GUT]
GRAVITATION decouples at 10^-43 seconds -
SUPERSUMMETRY
10^-45
3.16*10^19 GeV ,3.16*10^32K
BIG BANG

Figure 1.76.c. A pictorial description of various phase transitions during the expansion and cooling of the Universe.

[“The secrets of the Big Bang hide behind an impenetrable ball of fire”- The first split second, Govert Schilling, 31 st March 2001, New Scientist] .

Table 1.30. Unfolding of Events after Big Bang.

Theory applicable	Time after Big	Matter Density(gms/cc)	Temperature(Kelvin)	Energy(Mev)	Diameter of the Universe(cm)	The phase and

	Bang(seconds)	$\Omega = 1/(10^{-6}t^2)$ (gm/cc)[t is in seconds]	$T(K) = 10^{10}/[t(sec)]^{1/2}$	$T(MeV) = 1/[t(sec)]^{1/2}$	$R = 5\times10^{2}(cm/sec^{1/2})\times[t(sec)]^{1/2}$	phase transition
Theory of Quantum Gravity(yet to be developed)	0 to 10^{-43} seconds					Supersymmet
	Planck's time = $[Gh/c^5]^{1/2} = 10^{-43}$s	10^{93}gm/cc	10^{32}K	10^{19}GeV	Planck's Dia. = $[Gh/c^3]^{1/2} = 10^{-33}$cm	
	10^{-43}s to 10^{-35}s	Inflationary Phase(Fuelled by the energy released by phase transition)	Phase transition from supersymmetry to GUT[gravitational force decouples and relic gravitons are left out]			
On	10^{-34}s	10^{74}gm/	10^{27}K	10^{14}GeV	5×10^{-15}cm	Above

July 27, 1964, CP conservation violation detected by Chrswansten

cc

this temperature matter and anti-matter were produced in equal amount and annihilated. Therefore total matter was not conserved and Baryonic Number not conserved

Below this temperature, charge parity conservation was

						slightly violated in favour of matter and then onward baryonic number was conserved
10^{-10} s	10^{26} gm/cc	10^{15} K	100 GeV	5×10^{-3} cm	Salem-Weinberg Phase Transition	Electro-weak symmetry broken. Weak force decouples. Relic intermediate vector bosons are left out.
10^{-6} s	10^{18} gm/	10^{14} K	1 GeV	5×10^{-1} cm	Quark-	

cc

Nucleon Phase Transition

Strong force decouples and relic quarks left out. Abundance of quarks should be as high as gold. But no independent quarks detected till now.From this onward , quarks are always in

confined state either as triplet(baryons) or as doublet(gluons)						
	10^{-4}s	10^{14}gm/cc(nuclear density)	10^{13}K	100MeV	5cm	Electron neutrino decoupled.Mu neutrino decoupled.Ins of leptons leads to baryons instability.n← takes place)
	0.01s		10^{12}K			
	0.3s		1.8×10^{11}K			
	4s	$(10^{6}/16)$gm/cc	5×10^{9}K	0.5MeV	1000cm	Leptons stabilize.Tran stops.1n +7p ratio - stable equilibrium reached
	3					75%Hions

	minutes				+ 25%He ions
	300,000 yrs	10^{-20}(1.01) gm/cc	(10^{13} s)	10^{4}/3.3) = 4000K × 10^{8} cm; (10^{-6}/0.35) MeV = 0.3 eV	77%neutral H2 molecules + 23% neutral He atoms.
Electromagnetiuc forces decouple and relic radiation today is observed as cosmic microwave background radiation(2.7K CMBR)					
	13.7b.yrs	Today	2.77K		Present day Universe

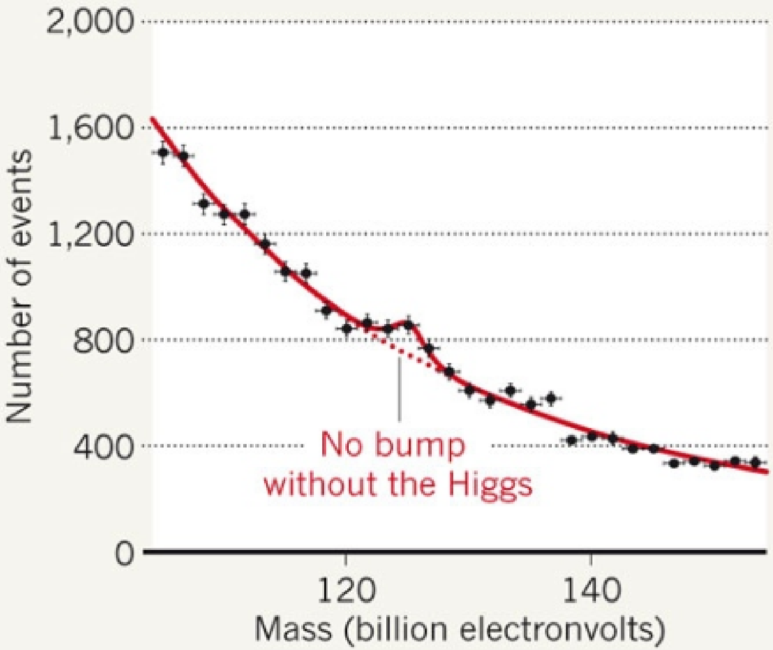
The Epic Quest has ended and the New Campaign has started

On 4th July 2012, the discovery of a new particle of mass 125GeV was announced which seemed to be consistent with Higg's Boson of the Standard Model.

The Epic Quest has ended but the new campaign has just begun.

BUMP OF DESTINY

The Higgs boson is most clearly seen by its decay into pairs of high-energy γ photons, indicating that it has a mass of around 125 gigaelectronvolts.



The greatest particle-physics discovery in a generation appeared as no more than a modest bump on a gently sloping plot (see ‘Bump of destiny’). Yet it drew a burst of applause as the two main experimental groups seeking the particle flashed their data onto the screen last week. The bump was the clear signal of a Higgs particle at a mass of around 125 gigaelectronvolts (mass and energy are used interchangeably in particle physics). Both the ATLAS and the Compact Muon Solenoid (CMS) detector groups reported that the significance of their signal was around five standard deviations — meaning that if the Higgs particle did not exist, there would be less than a one-in-a-million chance of getting these data by chance.

(Adapted from NATURE, Vol.487, Issue No.7406, 10th July 2012.)

SSPD_Chapter 1_Part 13_ Asymmetrical Universe After the Big Bang, matter and anti-matter were balanced.As Universe cooled down, at 10ˆ-10 sec, Salem-Weinberg Transition Phase decouples W and Z massive Bosons. These massive Bosons suppress certain transformations leading to CP violation. This CP violation leads to freezing out the baryonic and leptonic number. From then onward Baryonic and Leptonic number are conserved and from that our present Universe has evolved out.

SSPD_Chapter 1_Part 13_continued_Asymmetric Universe

1.13.4. WHY WE HAVE UNIVERSE MADE OF MATTER AND NOT ANTI_MATTER

Before Hubble's discovery (1929) of expanding Universe, it was sacredly held that matter is constant and conservation of matter holds good. Universe was suppose to be static and unchanging.

In 1928, P.A.M.Dirac successfully reconciled Quantum Mechanics and Special Theory of Relativity in his relativistic Wave Equation. This predicted the existence of anti-particle. It gave the correct value of the magnetic moment of electron. In May 1931, Dirac postulated an anti-particle in a following manner," a hole, if there were one, would be entirely a new kind of particle, unknown to experimental physics, having the same mass but opposite charge of electron.” This was a bold assertion. In 1932 the anti-particle of electron was discovered by C.D.Anderson in Cosmic Rays. This was positron. This was not proton. E.Seagre, O.Chamberlain and their colleagues produced anti-proton in High Energy Labs in 1955.

After that a series of anti-particle was discovered namely anti-proton , anti-neutron and mesons. It became difficult to classify mesons- are they particle or antiparticle. Now we know that they are admixtures of particle and anti-particle depending upon the composition of a doublet as already discussed in Chapter 1_Part13_constituent particles.. Rewriting Eq.1.125 we get

	Up quark	Strange antiquark	K+ Meson
Charge	(2/3) e_0	(1/3) e_0	+ e_0
Equivalent mass	310MeV	-505MeV	-195MeV
Spin	(1/2)ħ	(-1/2) ħ	0ħ

Through **C**harge conjugation, **P**arity inversion and **T**ime reversal we get the following anti-particle of positive K meson. To maintain CPT symmetry, existence of K+ Meson and Anti-K+ Meson or K- Meson is a physical reality.

	Up anti quark	Strange quark	Anti-K+ MesonOr K- Meson
Charge	(2/3) e_0	(1/3) e_0	e_0
Equivalent mass	-310MeV	505MeV	195MeV
Spin	(-1/2)ħ	(1/2) ħ	0ħ

These discoveries changed the basic law of conservation of matter. Today ant-particles and anti-atoms are being routinely created in High Energy Particle accelerators.

According to Dirac, there is a symmetry between particle and anti-particle. And this requires continuous creation and annihilation of matter and anti-matter.

e + anti e ↔ 2γ (soft gamma photons)

p + anti p ↔ 2γ (hard gamma photons)

$$\begin{array}{l} particle + anti\ particle \\ \leftrightarrow energy\ conserved \\ + total\ intrinsic\ angular\ momentum\ concerved \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ 1.135 \end{array}$$

After the Big Bang, both matter and anti-matter were created in equal amount. Somewhere down the path of evolution. Matter was favoured over anti-Matter. This led to the conservation of :

(Baryon-antiBaryon)/(no. of energy quanta) = Baryonic Number

And conservation of:

(Lepton- antiLepton)/(no. of energy quanta)=Leptonic Number.

Thus the Law of Conservation of Matter was replaced by “symmetry of matter and anti-matter” and “conservation of Baryon Number and Lepton Number”.

Implicit in this new conservation is that:

(Baryon- antiBaryon)= difference of the two kinds of Baryons is constant and

(Lepton-antiLepton)= difference of the two kinds of Lepton is constant.

But individually any of these four kinds can change.

A fundamental law of Physics is CPT symmetry. This implies symmetry of physical laws under transformation that involve conjugation of **C**harge, inversion of **P**arity or reversal of handedness and reversal of **T**ime simultaneously. CPT Theorem requires preservation of CPT symmetry by all physical phenomena.

The implication of CPT Theorem is that our Universe should have its mirror image known as Anti-Universe which should follow the same physical laws as are applicable to our present Universe..

Particle Physics should be invariant of CPT transformation. What this means is :

$$\begin{gathered}\textit{Description of Physics in an anti Universe (with opposite handedness and backward time)}\\ \equiv \textit{Description of Physics in Universe (with normal handedness and forward time)}\end{gathered}$$

In the following table we show how according to CPT Theorem, anti-particle is generated from particle.

	Matter particle							Anti-matter particle						
	m	Ji	τ	q	µm	Bn	Ln	m	Ji	τ	q	µm	Bn	Ln
baryon	m_0	J_0	τ_0	e	µ	1	0	-	$-J_0$	τ_0	-e	-µ	-1	0

baryon	m_0	J_0	τ_0	-e	-μ	--1	0	$-m_0$	$-J_0$	τ_0	--e	--μ	-1	0
baryon	m_0	J_0	τ_0	0	-μ	--1	0	$-m_0$	$-J_0$	τ_0	0	--μ	-1	0
gluon	m_0	J_0	τ_0	0	?	0	0	$-m_0$	$-J_0$	τ_0	0	?	0	0
gluon	m_0	J_0	τ_0	--e	?	0	0	$-m_0$	$-J_0$	τ_0	-e	?	0	0
gluon	m_0	J_0	τ_0	-e	?	0	0	$-m_0$	$-J_0$	τ_0	--e	?	0	0

m = mass, Ji = intrinsic angular momentum, τ = life time(in case of unstable atoms), q = charge , μm = magnetic moment, Bn = Baryon Number, Ln = Lepton number.

Neutron has negative magnetic moment (-1.91μN) and anti-neutron (+1.91μN) has positive magnetic moment. Though neutron has no charge but the constituent quarks have charges. These charges give rise to magnetic moment. The opposite arrangement in anti-particle gives rise to negative magnetic moment. In case of Neutron magnetic moment is negative hence anti-neutron has positive magnetic moment. Here μN refers to nuclear magneton.

In 1964 CP-symmetry violation was observed in neutral K meson decays.

Neutral Kaon decayed into anti-Neutral Kaon

And anti-Neutral Kaon decayed into Neutral Kaon.

These two decays rates should be identical to preserve CP symmetry. But the two decay rates were not identical. Neutral Kaon had a much larger life-time. This hinted at CP symmetry not being exact in a real Universe.

In 1967, Sakharov laid the foundation of "Asymmetrical Universe" more technically known as "Baryogenesis". He postulated the following:

1. Baryon number changing processes in the Early Universe only;
2. CP-symmetry violation;
3. An out of equilibrium situation for the Universe at time of baryogenesis.

What Sakharov was implying was that early in the Universe, Baryon and anti-Baryons were being equally created and annihilated so that under thermal equilibrium baryon number is equal to zero. But at some out of equilibrium situation, baryon number freezes out and then onward it is maintained constant.

By 1974-75, the three generation model of Particle became a reality.[3G model of Particle Physics became well established by 1992 as mentioned in Chapter1_Part 13_Big Bang] The third generation quark namely CHARM and third generation lepton namely Tau-lepton were discovered. With this

STANDARD MODEL of particle physics, CP-violation became its natural corollary.

Standard Model also predicts that baryon number changing processes (non-perturbative multiparticle effects) are very common at high temperature during the early part of Big-Bang but they become very rare later on at a much lower temperature of 3K which is CMBR temperature of present Universe. This is the reason why proton has a lifetime of 10^{32} years.

During the SALEM-WEINBERG transition, the out of equilibrium condition occurs for freezing out the baryon number. As already discussed in previous section at 10^{-10}second, Electro-weak symmetry is broken, weak force decouples, relic intermediate vector bosons are left out and from then onward baryon number and lepton number are conserved..

Demonstration of CP Symmetry.

Consider Parity Reversal operation. This operation is denoted by P. In this operation the object is MIRROR REFLECTED and ROTATED 180º as shown in Figure (1.77)

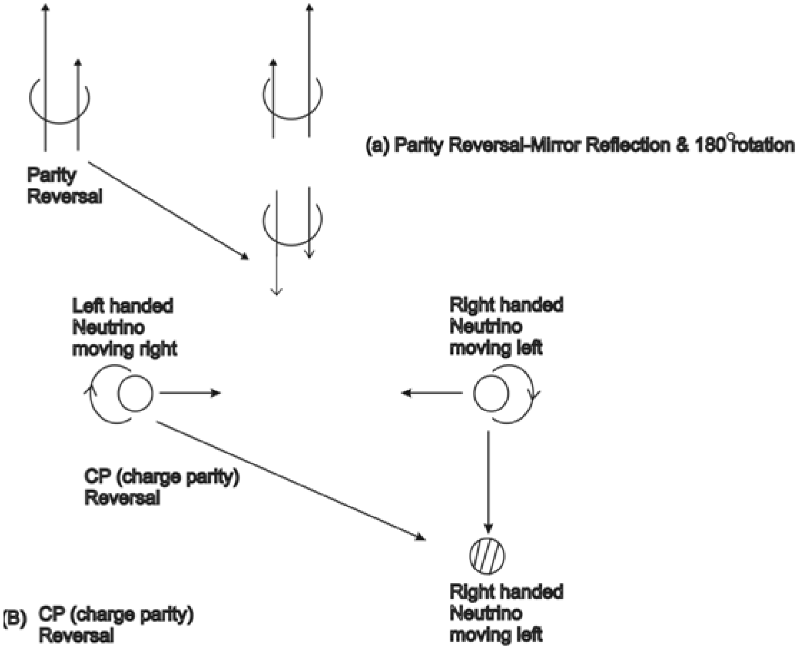

(a) Parity Reversal-Mirror Reflection & 180° rotation
Parity Reversal
Left handed Neutrino moving right
Right handed Neutrino moving left
CP (charge parity) Reversal
Right handed Neutrino moving left
(B) CP (charge parity) Reversal

Fig 1.77 Parity Reversal (Mirror Reflection and 180° Rotation) & Charge Reversal or Conjugation

Figure.1.77. Parity Reversal (Mirror Reflection and 180º Rotation) & Charge Reversal or Conjugation.

The laws of classical mechanics, electromagnetism and strong interactions are invariant under Charge-Parity (CP) reversal. But weak interactions violate CP reversal. That is the laws of weak interactions change under CP reversal.

Electro-magnetic forces are mediated by massless photons. Here parity is conserved. Photons do not distinguish between the handedness of electrons and other elementary particles.

Weak forces are mediated by heavy, electrically charged W bosons and heavy but electrically neutral Z bosons. W bosons interact only with left handed particles. Thus this causes maximum parity violation. Z bosons do not show maximum parity violation because of Weak Mixing Angle θ_W .

K-Mesons also known as Kaons and anti-Kaons decay at different rate. But this asymmetry is insufficient to explain the dominance of matter over anti-matter.

When electron and positron collide they produce massive B mesons. Stanford Linear Accelerator (SLAC), Stanford, USA, built BaBar and High Energy Accelerator Research Organisation(KEK), Tsukuba, Japan, built BELLE to study the decay rate of B mesons and anti- B mesons. The asymmetry in the two during CP reversal i.e. CP violation is measured by asymmetry parameter $\text{Sin}^2\,\theta_W$. This asymmetry ranges from -1 to +1. Zero value implies perfect symmetry under Charge Conjugation and Parity Reversal. It has been proved at confidence level of 99% that asymmetry parameter is 0.4 at both the asymmetry factories. This study remains inconclusive as to why matter has such a large dominance over anti-matter.

SSPD_Chapter 1_Part 14_BOSONS , FERMIONS and Physics of LASER ACTION.

Chapter 1_Part 14 from basic classical and quantum mechanics derives the basis of Fermi-Dirac and Bose-Einstein Statistics and leads to LASER Action.

SSPD_Chapter 1_Part 14_BOSONS , FERMIONS and Physics of LASER ACTION.

The essence of creation is formless and uniform. There is one law for all. This is the state of super symmetry. There is a quantum-gravity theory.

After the Big Bang as expansion takes place, space and time is created, energy density falls and temperature falls. Super symmetry is broken. Different energy particles and material particles are created.

Among these the fundamental particles are:

Intermediate Vector Bosons, leptons and lepton neutrino.

Hadrons are composite particles, composed of three generation of quarks and three generation of antiquarks. Triplets of quarks and anti-quarks are called Baryons and doublets of quarks and anti-quarks are called Mesons.

In 1910, Satyendra Nath Bose made an unique proposition while laying down the theoretical foundations of Black Body Radiation.

According to him all the integer number spin particles tend to aggregate together and these are called Bosons. In the energy spectrum their distribution function is defined as below:

P(ħω)=P(E)= 1/[Exp{(ħω)/(kT)} - 1] …………………………………… 1.136.

Where E = ħω ;

And P(E) = Probability of occupancy at the energy level E;

In contrast to this the particles which have half integer spin namely (1/2) ħ

or (3/2) ħ , they tend to lead solitary life and they are called Fermions. Their distribution function in the energy spectrum is the following:

P(E)= 1/[Exp{(E-E F)/(kT)} + 1] ……………………………………………. 1.137.

Where E F = Fermi Energy Level= the maximum energy level occupied by the particle in question at 0 K or where the Probability of Occupancy is (1/2) at any temperature T Kelvin.

Why do we encounter these two kinds of statistics or these two kinds of tendencies in nature: one of aggregation and the other of segregation?

If we look closely at these two particles we find a very significant feature between these two particles. Bosons are indistinguishable whereas Fermions are distinguishable.

1.14.1. THE EFFECT OF DISTINGUISH ABILITY ON DICE GAME.

Dice game is played as a game of gambling in almost all countries.

In its simplest form there are two coins each with two sides: head and tail. Therefore if they are played together and if the two coins are distinguishable then there are 4 distinct permutations:

Coin A- head, Coin B- tail;

Coin A- tail, Coin B- head;

Coin A- head, Coin B- head;

Coin A- tail, Coin B- tail.

But if the two particles are indistinguishable then we have only three combinations:

First combination: head and tail or tail and head;

Second combination: head and head;

Third combination: tail and tail.

When distinguishable coins are present then probability of occurrence of each of the four permutations is quarter(1/4).

When indistinghuishable coins are then probability of occurrence of each of the three combinations is one-third (1/3).

Thus indistinguishability has increased the probability of occurrence to:

(1/3)/(1/4)= (4/3) = 1.333

In the above example distinguishable coins are Fermions and indistinguishable are Bosons.

1.14.2. THE SCATTERING OF DISTINGHUISHABLE AND INDISTINGHUISH ABLE PARTICLES.

In Figure(1.78) the scattering system of two indistinguishable particles is shown. The indistinguishable particles are alpha particles.

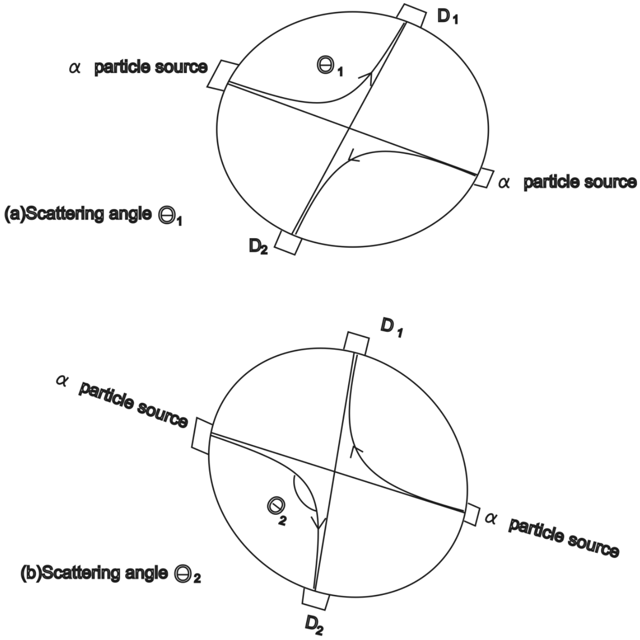


Fig 1.78 Scattering of two indistinguishable particles

Figure 1.78. Scattering of two indistinguishable particles.

We clearly see that an alpha particle is detected at a scattering angle θ_1 irrespective of the fact that the particle may be scattered at θ_1 or at θ_2 . This is because the alpha particles are indistinguishable. Hence

If probability amplitude of detection at θ_1 is $\psi(\theta_1)$

And probability amplitude of detection at θ_2 is $\psi(\theta_2)$, Then total probability density of detection of alpha particles at θ_1 is

$| \psi(\theta_1) + \psi(\theta_2)|^2$ **1.138**

In the same scattering set up if α- α pair is replaced by α-O_2 pair the probability density of detection of α particles at θ_1 is

$| \psi(\theta_1) - \psi(\theta_2)|^2$ **1.139**

This is because the scattering at θ_2 gives the detection of O_2 which is completely distinguishable from the alpha particle.

Eq.(1.138) and Eq.(1.139) bring out the fundamental difference between bosons and fermions.

1.14.2. SCATTERING OF N PHOTONS LOCALIZED IN THE SAME SPATIAL CELL.

In one elemental phase cell (meaning by at the same energy level and in the same spatial space) only two fermions of opposite spin are accommodated. If a third fermion pushes itself in the cell it will be repulsed. Therefore we say fermions are segregative.

But in an elemental phase cell any number of photons can aggregate. As the number increases the tendency to aggregate increases. A quantitative measure of aggregation tendency comes out while theoretically analyzing Black body Radiation.

Suppose in a cavity the emission and absorption of photons are in equilibrium then the average energy at a given frequency is shown to be:

$$\langle E \rangle = \hbar\omega/[\mathrm{Exp}(\hbar\omega/(kT)) - 1] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ 1.140$$

Planck assumed that every black body radiator is composed of harmonic oscillators which oscillate at fundamental and harmonic frequencies only

that is at ω, 2 ω, 3 ω, 4 ω….. and their discrete energies are $\hbar\omega$, 2 $\hbar\omega$, 3 $\hbar\omega$, 4 $\hbar\omega$..

If total number of oscillators is N_{00} and if we assume that they are in energy equilibrium then from Maxwell – Boltzmann statistics:

If N_0 is the number of oscillators at 0 energy then

$N_0/ N_{00} = \mathrm{Exp}(-0/kT) = 1$

N_1 is the number of oscillators at $\hbar\omega$ then

$N_1/ N_{00} = \mathrm{Exp}[-\hbar\omega/kT]$

N_N is the number of oscillators at $n\hbar\omega$ then

$N_N / N_{00} = \mathrm{Exp}[-N\,\hbar\omega/kT]$ ……………………………………………………………
1.141

Equation(1.141) have been set up according to Maxwell-Boltzmann Statistics which states that at temperature T Kelvin, probability of existence at energy level E is :

$P(E) = \mathrm{Exp}(-E/kT)$ ………………………………………………………………
1.142

So the average energy (per oscillator) = total energy/ total number of oscillators

$<E> = E_{total}/N_{tot}$

$= [(0\times N_0) + (\hbar\omega\times N_1) + (2\hbar\omega\times N_1) + (3\hbar\omega\times N_3) + \ldots..]/[N_0 + N_1 + N_2 + N_3 + ..]$ ……………………………………………………………………
1.143

Dividing Numerator and Denominator by $N_{tot} = N_{00}$,

We get:

$<E> = E_{total}/N_{tot}$

$= \hbar\omega\,[(0\times N_0)/N_{00} + (1\times N_1)/N_{00} + (2\times N_1)/$

N_{00} + $(3\times N_3)/N_{00}+.....]/[(N_0)/N_{00}+(N_1)/N_{00}+(N_2)/N_{00}+(N_3)/N_{00}+..]$

assuming $N_N/N_{00} = x_N=$ Exp[- N ħω/kT] where $x=$ Exp[- ħω/kT] and substituting in Eq.(1.143), we get:

$<E> =$ ħω$[0+x+2x_2+3x_3+...]/[1+x+x_2+x_3+...]$ **1.144**

But $[1+x+x_2+x_3+...]$ is a Geometric Progression Series where x« 1 therefore the sum of this G.P. series is 1/(1-x)

Hence denominator of Eq.(1.144) is 1/(1-x) 1.145

Numerator $= [x+2x_2+3x_3+...]$

$= x[1+x+x_2+x_3+x_4+x_5+x_6+..........$

$+x+x_2+x_3+x_4+x_5+x_6+...........$

$x_2+x_3+x_4+x_5+x_6+$**..........**

$x_3+x_4+x_5+x_6+............$

$=x[(1/(1-x)+x(1/(1-x))+x_2(1/(1-x))+x_3+.........]$

Numerator =x(1/(1-x)) [1/(1-x)] ....

1.146

Substituting Eq.(1.145) and Eq.(1.146) in Eq.(1.144)

$<E> = \hbar\omega \{x(1/(1-x)) [1/(1-x)]\}/ 1/(1-x)$

$= \hbar\omega\, \mathrm{Exp}[- \hbar\omega/kT]/[1- \mathrm{Exp}[- \hbar\omega/kT]]$

**$<E> = \hbar\omega/ [\mathrm{Exp}[\hbar\omega/kT] -1]$ ....
1.147**

**At $\hbar\omega$, average energy = $<E>$ = $<n>$ $\hbar\omega$...
1.148**

Comparing Eq.(1.147) and Eq.(1.148):

**$<n>$ = average number of oscillators at $\hbar\omega$ $=1/[\mathrm{Exp}[\hbar\omega/kT] -1]$ ....
1.149**

Dividing numerator and denominator by Exp[ħω/ kT]

**$<n> = \mathrm{Exp}[- \hbar\omega/kT] / [1- \mathrm{Exp}[- \hbar\omega/kT]$ ....
1.150**

or **$<n> + 1 = 1/[1-\mathrm{Exp}[- \hbar\omega/kT]$ ....**

1.151

Therefore dividing Eq.(1.150) by Eq.(1.151)

$<n>/[<n>+1] = \mathrm{Exp}[-\hbar\omega/kT]$ ………………………………………………………………… **1.152**

From Statistical Mechanics under energy equilibrium:

$N_e/N_g = \mathrm{Exp}[-\Delta E/kT] = \mathrm{Exp}[-\hbar\omega/kT]$ …………………………………………………………………… **1.153**

Where N_e = number of atoms in excited state,

N_g = number of atoms in ground state ;

In energy equilibrium number of photons emitted is equal to number of photons absorbed.

By comparing Eq.(1.152) with Eq(1.153) we obtain:

$<n>/[<n>+1] = N_e/N_g$

Therefore **$N_g <n> = N_e [<n>+1]$** …………………………………………………………… **1.154**

If we assume that the probability of emission of photons is a_2 when there are no photons in the cavity then

$N_g <n> a_2 = N_e [<n> + 1] a_2$ **1.155**

At a given temperature T Kelvin, $<n>$ is the average number of oscillators at a circular frequency ω.

Since ground state atoms will absorb photons and excited state atoms will emit photons hence:

$N_g <n> a_2$ = rate of absorption

$N_e [<n> + 1] a_2$ = rate of emission.. **1.156**

The two rates are exactly balanced as is evident from Eq.(1.155).

Eq.(1.156) tells us that if $<n>$ number of oscillators are at the same frequency ω

then probability of absorption is $<n> a_2$ and

probability of emission is $[<n> + 1] a_2. = [<n> a_2 + a_2]$

Probability of emission = Probability of stimulated emission + Probability of spontaneous emission;

Here stimulated emission is induced emission. A passing photon induces an excited atom to emit a photon of the same frequency and same phase and in the process settle to ground state. The photon emitted by this process of stimulated emission adds to the existing photon. This is known as **L**ight **A**mplification by **S**timulated **E**mission **R**adiation. The acronym of this process is LASER action.

Spontaneous Emission takes place when an excited atom settles down to ground state by itself according to Life-Time Law.

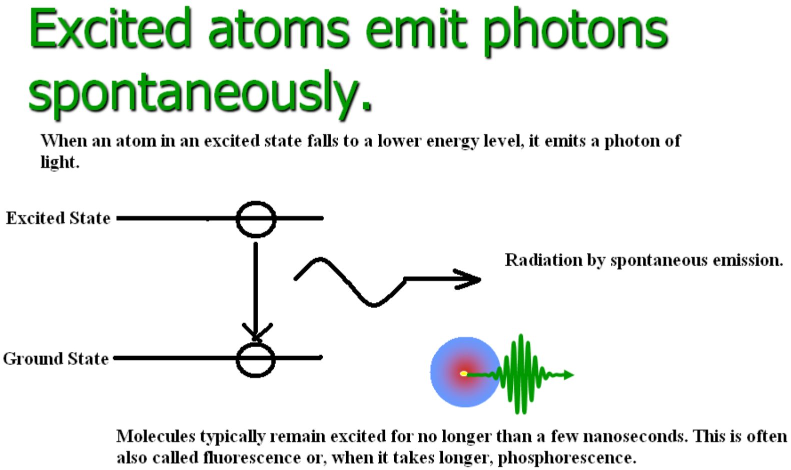
Einstein believed that stimulated emission is proportional to the intensity of light causing stimulation and intensity of light is given by the number of photons involved per second.

Therefore stimulated emission multiplicative factor

= absorption multiplicative factor $= <n>\ a_2$

spontaneous emission multiplicative factor $= a_2$
…………………………………………………… **1.157.**

This bosonian property of stimulated emission is utilized in LASER, the acronym for **L**ight **A**mplification by **S**timulated **E**mission **R**adiation. In Figure 1.79, the physics of spontaneous emission is shown.



Excited atoms emit photons spontaneously.

When an atom in an excited state falls to a lower energy level, it emits a photon of light.

Molecules typically remain excited for no longer than a few nanoseconds. This is often also called fluorescence or, when it takes longer, phosphorescence.

Figure 1.79. Radiation by Spontaneous Emission.

When an atom relaxes from excited state to ground state in a natural way then we refer to it as spontaneous emission. If an excited solid experiences relaxation it gives out an incoherent beam of light.

A solid can similarly be brought to excited state by irradiation by an incoherent beam of light with a frequency greater then its threshold frequency. The physics of irradiation and subsequent excited state is shown in Figure 1.80.

Atoms and molecules can also absorb photons, making a transition from a lower level to a more excited one.

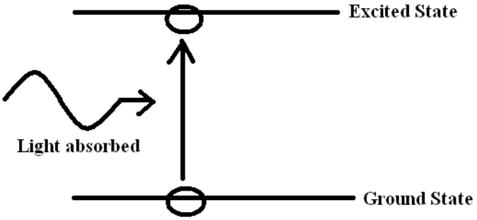


Figure 1.80. An atom absorbs light of the correct frequency and gets raised to an excited state.

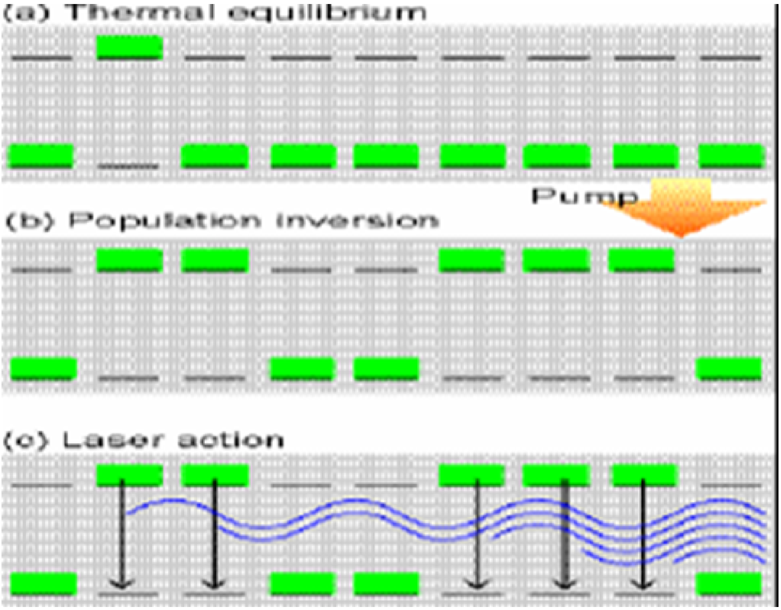


Figure 1.81. Through light irradiation RUBY is being brought to excited state. The spontaneous emission kick starts the LASER Action. The photon emitted by spontaneous emission causes stimulated emission. By stimulated emission a photon of same frequency and phase is emitted. Now there are two electrons.

Two electrons are in phase and of same frequency and they constitute a coherent wave. This coherent wave causes two more stimulated emission. This results in 4 photon strong coherent beam. In this way the coherent beam exponentially multiplies and becomes strong enough to emerge out of one of the partially reflecting surfaces. This is known as Light Amplification by Stimulated Emission Radiation.

In a LASER device first energy from DC power source is utilized to cause the population inversion. All the ground state atoms are brought to an excited metastable state through irradiation by suitable light source. Metastable means that these atoms will remain in excited state for periods of milliseconds and then if no stimulation takes place they will spontaneously settle down to ground state radiating off photons as incoherent light.

In a LASER the first photon is obtained through spontaneous emission. This photon causes stimulated emission of in-phase and same optical frequency photon.

The two photons become 4 photon in-phase , same frequency beam. When a transition to lasing mode occurs, all atoms cooperate and emit in synchronism- producing giant coherent wave. A spontaneous emission changes into stimulated emission. A chaotic system self organizes itself into an orderly system. This beam geometrically grows into a powerful coherent beam which can cause miracles. This is illustrated in Figure 1.81.

SSPD_Chapter 1_Part 15_Quantum Fluids_Superconductivity & Superfluidity.
SSPD_Chapter 1_Part 15 describes the properties of solids below degeneracy temperature.

SSPD_Chapter 1_Part 15_ SUPER-FLUIDITY AND SUPERCONDUCTIVITY-A SUBCLASS OF QUANTUM FLUIDS.

[Electrical Engineering Materials by A. J. dekkar, Publisher PHI]

[The New Physics, edited by Pau; Davies, ‘Low Temperature Physics, superconductivity and superfluidity’ by Anthony Leggett, pp 268, Cambridge University Press, 1992 .]

We saw in Chapter 1_Part 3_Section 1.3.1 that below degeneracy temperature quantum mechanical effects are markedly evident. For solids and liquids, the atoms wave nature is exhibited below the temperature

$T_{degeneracy} = h^2/(3kM \times ((m_p + m_n)/2) \times (a \times 10^{-10} m)^2)$

$= 633.3/(M \times a^2)$ Kelvin

where M is the mass number of the atom and ‘a’ is the distance between the scatterer in Angstrom. In order to observe super fluid and super conductors we must go to low enough temperatures to obtain quantum liquid. The following two conditions have to be satisfied to obtain quantum liquids:

1. cold enough to be treated as degenerate systems

that is the liquid should be below degeneracy temperature;

(ii) it should be a bosonic system.

An assembly of odd number of fermions is fermionic and an assembly of even number of fermions is bosonic.

In classical mechanics particles follow Maxwell-Boltzman statistics . In a large number of gas particles system of volume V and gas molecules number N, each particle has a velocity v and energy E = Kinetic Energy + Potential Energy = $[(1/2)mv^2 + 0]$ .Here we assume that Potential Energy is zero. Maxwell- Boltzman distribution tells that number of particles at energy E is n(E) where ;

$n(E) = K \times \mathrm{Exp}[-E/kT] = K \times \mathrm{Exp}[-mv^2/(2kT)]$ ……………………………….**1.158**

In a quantum mechanical system kinetic energy will be quantized just as the energy of electron in a potential well is quantized corresponding to the standing waves of matter wave of different quantum numbers. But here the well corresponds to the large volume hence quantum difference between the discrete energy states is miniscule therefore discrete energy states E_i can be treated as a continuous variable while plotting the graph. In this graph plot, n_i is the number per quantum state whereas in classical systems n is the number per unit energy .

Therefore $n_i(E_i) = K \times \mathrm{Exp}[-E_i/kT]$ ……………………………………………………………… **1.159**

In quantum mechanical systems, we have density of permissible quantum states N(E) per unit energy.

Hence $n_i(E_i) = N(E_i)P(E_i)dE$ = number of particles lying between E_i and $(E_i + dE)$

Where $N(E_i).d(E)$ = number of permissible quantum states between E_i and $(E_i + dE)$

And $P(E_i)$ = probability of occupancy at $E_i = \mathrm{Exp}[-E_i/kT]$ according to Maxwell- Boltzman Statistics ;

Therefore Equation(1.159) can be written as :

$n_i(E_i) = N(E_i)dE \times \mathrm{Exp}[-E_i/kT]$ …………………………………………………… **1.160**

Plot of Equation (1.159) is given in Figure (1.82). The plot of Equation(1.160) will require $N(E_i)dE$to be accounted for . $N(E_i)dE$is not a constant but a function of E_i

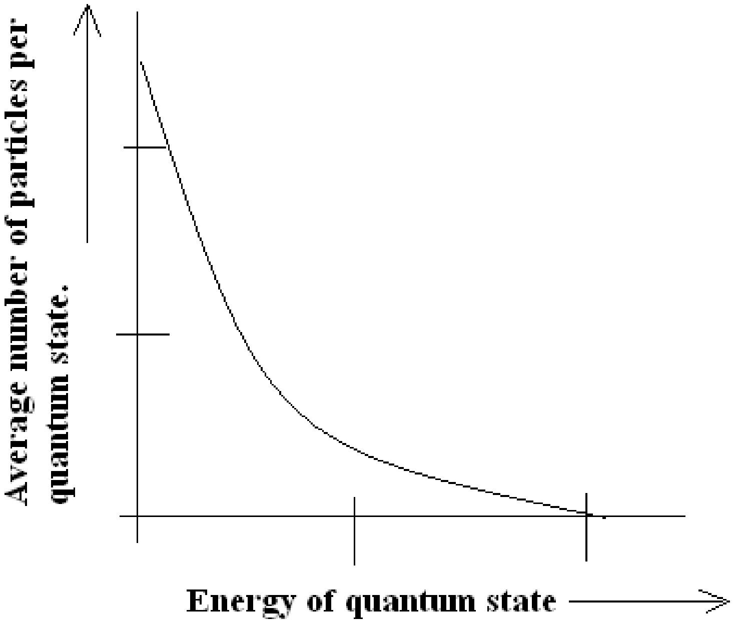


Figure 1.82. The average number of particles per state at a given temperature for particles obeying Classical Statistics.

The average energy per particle comes out to be (3/2)kT for the classical distribution as given by Equation (1.158). The average energy obtained by Equation (1.159) which applies to a system of distinguishable particles is practically the same except at very low temperatures 10^{-11}Kelvin which is of no practical consequence in macroscopic systems.

But a system of distinguishable particles follows Fermi-Dirac Statistics as shown in Chapter 1_Part 8_Section1.8.2.3. At low temperatures it has the form shown in Figure (1.83). At 0 Kelvin there are 2 electrons or 2 fermions of opposite spins per quantum state up to EF (Fermi Energy) and above EF the permissible states are empty. This rectangular distribution becomes more and more skewed with rise in temperature until above degeneracy temperature the system of distinguishable particles becomes non- degenerate and behaves classically as shown in Figure (1.82).

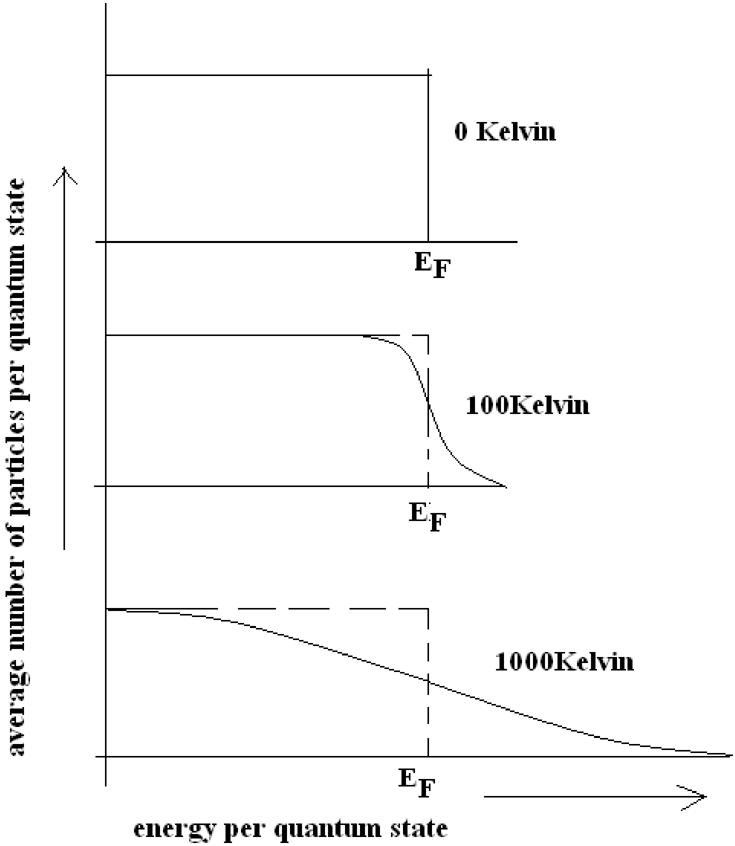


Figure 1.83. The average number of particles per permissible quantum state for particles obeying Fermi-Dirac Statistics at Temperature (a) 0 Kelvin (b) T 1 Kelvin (c) T 2 Kelvin.

A very illustrative example of these two conditions are the conducting electrons in a metal which is a degenerate system following the distribution in Figure (1.83) and conducting electrons in semiconductors which is non-degenerate following Figure(1.82).

Here it is noteworthy that in Figure (1.82) the number of electrons is 1 or less. Here 1 means 1 pair of opposite spin. Less than 1 means the state has less than 100% occupancy.

For Bosonic system consisting of indistinguishable particles things are very different. As shown in Figure (1.84), The number of particles at lower energies is much higher than the classical case If we have a fixed number of particles then at low temperature we have Bose Condensate which we have talked about in Chapter 1_Part 8_1.8.2,2. At higher temperature the distribution of Bosons cannot be distinguished from that of Fermions.

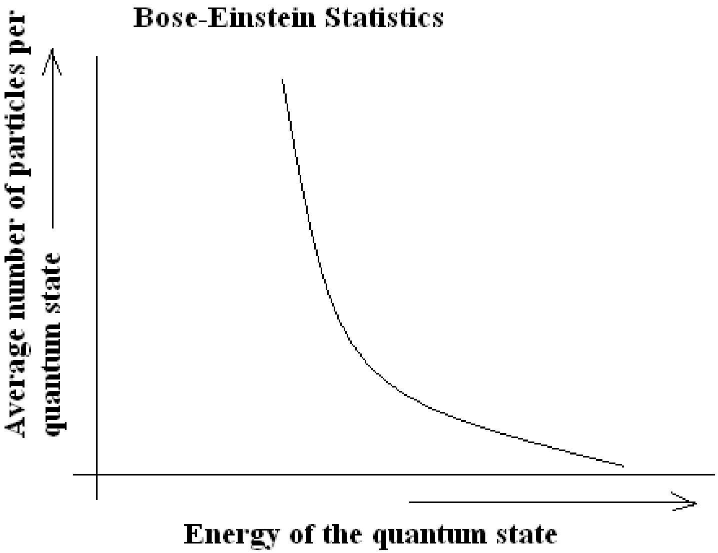


Figure 1.84. The average number of particles per state at a given temperature for particles obeying

Bose-Einstein Statistics.

This discussion implies that at low temperatures the distribution is going to be critically effected by the fact that the system is bosonic or fermionic.

In solids we do not see any difference in the distribution in Bosonic System say Diamond isotope 12 (6 p’s, 6 n’s and 6 e’s) and Fermionic System say Diamond isotope 13 (6 p’s, 7 n’s and 6 e’s). Whatever difference is seen can be explained in terms of atomic masses. Here Diamond isotope 12 though indistinguishable are still distinguishable since the individual atoms are tied to their respective positions. The difference in the two distribution shows up in liquids and gases.

In gases the density is very low hence interatomic distances is very large and so the degeneracy temperature is too low for observing the difference.

So to see the difference in a bosonic system and fermionic system we need a quantum liquid. It is sufficiently dense so that we have an achievable degeneracy temperature . At the same time it is sufficiently mobile so that an indistinguishable system really becomes indistinguishable. This liquid should be at a temperature below Degeneracy Temperature. Then it is a proper quantum liquid.

Excepting element He, all elements and compounds freeze at degeneracy temperature T_0 . Helium remains liquid all the way to 0 Kelvin so it is an appropriate candidate for the study of super-fluids. This good fortune is due to low atomic mass and weak inter-atomic forces. Weak inter-atomic force is due to the completion of octave in the outer filled orbit. It does not freeze under its own vapor pressure but it does solidify if the external pressure is increased to 30 atmospheric pressures. Also luckily we have two stable isotopes 4He and 3He providing bosonic system and fermionic systems respectively. 4He has 2 p’s, 2 n’s and 2’s and therefore it forms a bosonic system. 3He has 2p’s, 1n and 2’s (produced in beta decay of tritium from nuclear reactors) and therefore it forms a fermionic system. These two liquid systems at temperatures below degeneracy temperature provide an ideal laboratory material for the study of quantum statistics (Fermi-Dirac/Bose-Einstein).

The electrons in solid metal or liquid metal are degenerate systems since, as we saw in Section 1.3.1., the degeneracy temperature is of the order of 10,000Kelvin.

The electrons in White Dwarf and neutrons in Neutron Stars are degenerate systems. Because of very high density in these compact stars, the degeneracy temperature is very high of the order of 10^{10} K. Nevertheless they form quantum liquids.

In Table (1.32) tabulates the quantum liquids, their degeneracy temperatures and the onset temperature of superfluidity/superconductivity.

The charged systems containing conducting electrons exhibit superconductivity and neutral atomic systems exhibit super-fluidity at temperatures much below degeneracy temperatures.

Table 1.32. Laboratory Quantum Liquids. (Table 9.1. Low Temperature Physics, super conductivity and super fluidity. By Anthony Leggett, ‘The New Physics’ edited by Paul Davies, Cambridge University Press, 1992.)

System	Statistics	Degeneracy Temperature(K)	Superconductivity/ Superfluidity?	Onset Temperature(K)
Electrons inMetals	Fermi	**~ 10 ^4**	Sometimes	< 125
Liquid He 4	Bose	~ 3	Yes	2.17
Liquid He 3	Fermi	~ 5	Yes	**2.6× ×10 ^-3**
Atomic H	Bose	**~ 5×10 ^-2**	?	?
Atomic D(deuterium)	Fermi	**~ 5×10 ^-2**	?	?

1.15.1. LIQUID HELIUM-4 AS A SUPERFLUID.

Helium $_2He_4$ has two protons, 2 neutrons and 2 orbital electrons. Since all the particles are in pairs hence total spin angular momentum is zero as well as electrons orbital angular momentum is zero. Therefore two Helium atoms are indistinguishable. In this sense Helium atom is a boson.

Above 4 Kelvin, Helium is in gaseous state. At 4 K it liquefies. This liquid does not have any special property. It exhibits the same properties as any other viscous liquids. This is called He-I phase.

As we cool it below $T_\lambda = 2.17$ K there is a phase transition to He-II and in He-II phase it exhibits abnormal and spectacular behavior such as :

1. it flows through tiny capillaries without experiencing any friction;
2. it climbs in the form of film over the edge of vessels containing it. This phenomena is referred to as ‘film creep’;
3. spouts in a spectacular way when heated under certain conditions;
4. if contained in a rotating container, the content in He-II phase never rotates along with the container;

This complex of effects is called SUPERFLUIDITY and the liquid is in superfluid phase which we referred to as He-II phase. The transition between two phases is called λ transition and this occurs because of the onset of Bose Condensation.

1.15.2. What is BOSE CONDENSATION ?

Let us begin with He-I phase with N atoms in volume V at a temperature T below 4 Kelvin. According to Bose Statistics N is distributed as N_1 in low states and N_2 in high states. As the temperature is lowered more and more He-4 is transferred to lower states so that $N_1 + N_2 = N$.

This transfer stops when Nreaches N_{max} corresponding to the lower energy state at a lower temperature because there is an upper limit of the number of quantum states available for the given energy state. The temperature at which this occurs happens to be the degeneracy temperature T_0 which we referred to in Section (1.13.1). This is the λ phase transition with He-I transforming into He-II.

At a temperature $T^* < T_0$, $N_{max} < N$ because N_{max} is a function of temperature and it reduces as temperature is decreased. Now Bose distribution cannot accommodate N . So how is the excess atoms accommodated that is excess over N_{max} . Nature has an ingenious method to circumvent this problem- **BOSE CONDENSATION**. The excess amount at T* condenses to the ground state. As T* approaches

ZERO Kelvin absolute temperature, all the N atoms are condensed to the ground energy state.

This large scale transfer of He-II to energy ground state is what gives rise to super-fluids and its spectacular property of super-fluidity as mentioned in Section (1.15.1).

The degeneracy temperature of He-4 is calculated to be 3 K and the onset of super-fluids is 2.17K. This discrepancy occurs because Degeneracy Temperature is calculated for systems with non-interacting particles whereas as He-4 is strongly interacting.

1.15.3. SUPERCONDUCTIVITY.

In 1911 Kamerlingh Onnes discovered that liquid mercury at temperatures below 4.2Kelvin shows zero resistance. This was christened as super conductivity. There is a very narrow temperature range in which the transition takes place. This is defined as Transition Temperature T_C.

In 1986, J.G Bednorz and K. A. Muller, workin.g at IBM Research Laboratories at Zurich, found a new class of materials which are superconducting at a much higher temperature. Yitrium Barium Copper Oxide (YBaCuO) was discovered to behave as superconductor at 77Kelvin (liquid Nitrogen temperature). This started the era of high temperature ceramic superconductors.

Presently super-conductors have been achieved till 125K.

Professor Collins induced a current to flow in superconducting lead ring. This continued to flow for a year. This confirms that indeed resistance is zero in a super conductor.

In 1933, Meissener and Ochsenfeld discovered that a superconductor is a perfect diamagnet. This implies that magnetic flux density B in a superconductor is zero. This has been called Meissener Effect. If a coin size superconductor is placed on one of the magnetic poles then it will be freely suspended in air because of the repulsion force between magnetic pole and the diamagnetic nature of the coin size superconductor. This is known as the magnetic levitation effect.

At low temperatures, ordered behavior appears spontaneously in a macroscopic system. Because of this ordered behavior, the given system exhibits super fluidity and super conductivity. Enormous number of particles behave coherently to give rise to frictionless flow of liquids (super fluidity) and frictionless flow of electrons(super conductivity).

A kind of phase transition occurs just as there is phase transition from gas to liquid to solid or the onset of ferromagnetism in iron when it cools below the Curie temperature. In the same way lowering of temperature causes electron pairing into cooper pairs. These cooper pairs suppress the fermionic properties to produce super conducting state. This is spontaneous self organization.

1.15.3.1. PHYSICS OF SUPERCONDUCTIVITY.

A metal at room temperature has a metallic bonding. As we discussed in Section (1.12) the metallic atoms release their outer orbital electrons at room temperature. These semi-free electrons (semi-free because they are free to move around the whole lattice but confined within the boundaries of the crystal because of the surface barrier potential $\varphi_{SB} = W_F$(Work Function of the metal)/q) belong to the whole metallic lattice and not just to their parent atoms. This conducting electron gas hold the regular crystalline array of positively charged ionic cores together as a metallic solid.. In alloys these ionic cores are held irregularly. This electron gas follows Fermi- Dirac Statistics as shown in Figure (1.83).

At zero Kelvin all the quantum states are filled up to Fermi Energy level (E_F) with a pair of opposite spin electrons in each quantum state. Above E_F all the states are empty. But as the temperature is raised the distribution becomes skewed and the upper electrons within kT energy segment start jumping to higher energy states. This is the reason why the actual specific heat of the electron gas is kT/E_F of the classically expected value.

By taking the effective mass of the conducting electrons into account we can describe these electrons by de Broglie matter waves traveling in free space. The effective mass which may range from 0.1m_e to 10m_e accounts for the deformation caused in the de Broglie free wave. But the crystal defects, foreign impurities and the thermal vibration of the ideal regular array of ionic cores are there to scatter the matter wave. These factors limit the mean free path to several inter atomic spacings. This is the reason for the limited mobility , limited conductivity and finite resistance. But if the thermal vibrations are frozen out by cryogenic cooling to 4K (liquid He-4 temperature) , impurities are absent and the crystal is defect-free then the mean free path is of the order of centimeter and mobility is drastically improved and resistance fall to a residual value. This ideal gas model is correct by and large and is able to explain the general features of the behavior of metal.

But there is a class of metals and ceramic conductors which experience a definite phase transition at a critical temperature or transition temperature T_C just as we saw λ- phase transition in He-4. In He-4 below T_0 (the degeneracy temperature) the bosonic Helium-4 atoms start condensing to energy ground state and manifest the behavior of a super-fluid.

In metals above Tc the electron gas is a fermionic system. At temperatures below Tc the fermionic electron gas is abruptly condensed into an assembly of boson molecules called cooper pairs of electrons. These cooper pairs experience BOSON CONDENSATION in an analogous manner as in He-II phase and manifest a complex of spectacular properties characteristic of a super-conductor namely:

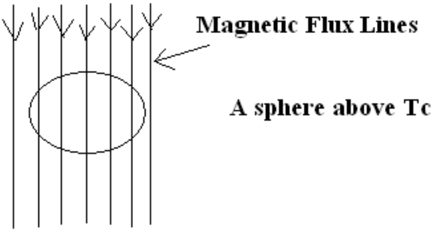
1. the electrical resistance becomes zero. If a super conducting ring is cooled below the transition temperature and magnetic field linking the coil is varied for a short period of time then a current is induced and it continues to flow in the super-conducting ring for time immemorial. In super-conductors the usual scattering mechanisms are rendered ineffective just as in He-II the frictional effect of the walls of the container are rendered inoperative;
2. ‘Meissner Effect’- that is the total exclusion of magnetic flux from a super conducting body. By Lenz’s law of electromagnetic induction, electro magnetic force (e.m.f.) induced in a conductor is proportional to the rate of change of magnetic flux linking the conductor and the direction of the e.m.f. and the consequent eddy current is such as to oppose the cause of change. If the change of flux is increasing the eddy current will produce an opposing magnetic field and if the change of flux is

decreasing then the eddy current will produce an aiding magnetic field. In effect the eddy current tends to shield the conductor from a change of flux. In ordinary conductors the eddy current ebbs out in seconds and the shielding effect vanishes in the same time scale and the magnetic field assumes the new steady state value of the magnetic field. But in astrophysical objects the eddy current and the shielding effect persists for several thousand years. In super-conductor since the resistance is zero, the eddy current persists for ever and continues to exclude the magnetic flux for time immemorial. Effectively a superconductor behaves as a dia-magnetic material;

3. Magnetic Levitation- this is a direct consequence of Meissner Effect. In the process of excluding the magnetic field the superconductor experiences a force of repulsion. If a superconductor is held over the north pole or south pole of a magnet, the former will be held hanging in the mid-air because the gravitational force on the freely falling body is exactly counter-balanced by the force of repulsion due to field exclusion;
4. Magnetic Flux Quantization- If a ring of superconductor is placed in a magnetic field then the field excluded from the annular part of the ring but is included through the central hole of the ring. The flux linking the central hole is an integral multiple of $h/e = 10^{-15}$

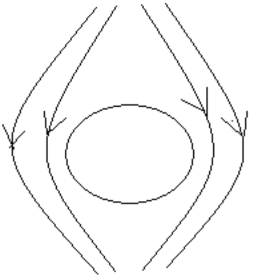
Weber i.e. φ = n(h/e) where n is an integral number. The flux passage through a normal non-magnetic conductor, the magnetic flux exclusion by the superconductor and the partial linkage through a super-conductor ring is shown in Figure (1.85).

Fig 1.85.a



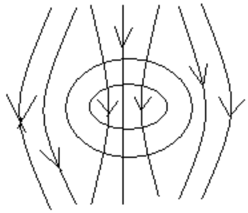
A sphere above Tc

Fig 1.85.b



A sphere below Tc hence it becomes superconductor and excludes magnetic flux line from its interior

Fig 1.85.c



A toroidal superconductor keeping out the magnetic flux line but including Magnetic flux line within the space inside the toroidal ring.

Figure 1.85. Behaviour of a piece of superconducting metal in a magnetic field.

1. Above the transition temperature when the metal is normal and the sample is completely

penetrated by the field (whatever its shape is),
2. **Below T c when it is super conducting, a spherical sample completely excludes the field,**
3. For a ring shaped specimen in the super-conducting state, some of the field lines can go through the central hole of the ring but always as an integral multiple of (h/e).

The theory of cooper pairs and the formation of cooper pairs in isotropic state explains only ‘Very low temperature’ super conductors. There is a class of heavy fermion superconductors as well as high temperature ceramic superconductors which cannot be explained by the present theory.

SSPD_Chapter 1_Part 16_FERMIONS_Part 17_Density of States in Conduction Band of Metals
Chapter 1_Part 16 and Part 17 describes the Fermi-Dirac Statistics and derives the density of permissible quantum states in Conduction Band of Metals in Bulk Devices.

SSPD_Chapter 1_ Part 16 + Part 17.

1.16. FERMION AND FERMI DIRAC STATISTICS.

In nature, particles with $(n+1/2)\hbar$ spin angular momentum are distinguishable particles and hence a pair of particles with opposite spin occupy an elemental phase space and no more. Hence at one time, two particles will occupy different energy states or at one spatial coordinate two particles will be at different momenta.

These particles obey Heisenberg's Uncertainty Principle:

$\Delta E.\Delta t \geq \hbar$ 1.161

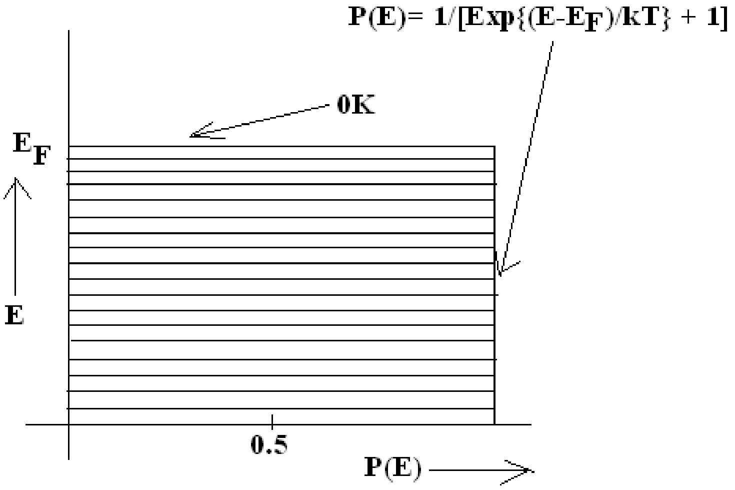
$\boldsymbol{\Delta p\ x\ .\Delta x \geq \hbar}$

Heisenberg's Uncertainty Principle leads us to Pauli Exclusion Principle. Pauli Exclusion Principle directly comes from Quantum Mechanics. Until the advent of Quantum Mechanics we could not arrive at the correct definition of Specefic Heat.

According to Quantum Mechanics, fermions obey the following distribution function also known as Fermi-Dirac Statistcs:

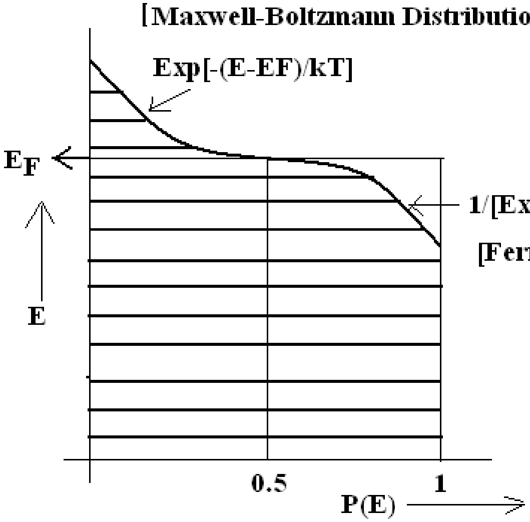
$$\mathbf{P(E) = 1/[\,Exp\{(E-E_F)/(kT)\} + 1\,] \quad ....1.162}$$

This distribution function or probability function tells us that at 0 Kelvin absolute temperature the probability of occupancy till E_F energy level is Unity and above E_F probability of occupancy is zero. This implies that energy levels till E_F are occupied and above are empty. Each level till E_F are occupied by two oppositely spinning electrons. At 0 Kelvin the probability distribution function is rectangular as shown in Figure(1.86).



Figure(1.86) At 0 Kelvin, fermions having rectangular probability distribution function.

At T Kelvin, probability distribution is skewed and the probability of occupancy at EF is 50% . This means that half of the time EF is occupied and half of the time it is empty and electrons are occupying a higher energy state. The probability of occupancy distribution at T Kelvin is shown in Figure(1.87).



Figure(1.87) Skewed Probability Distribution of Fermions at T Kelvin.

As shown in Figure(1.87), around Fermi-level EF and below number of electrons, n, is only slightly less than the available allowed energy states N(E)dE. This is known as DEGENERATE STATES and fermions follow Fermi-Dirac Statistics.

At energies E- EF > > kT and at 300 Kelvin kT=

0.026eV that is at E equal or greater than (E_F + 0.26eV), the distribution function is the following:

P(E) = Exp [-(E- E F)/(kT)] 1.163

The above distribution function is the classical distribution function known as Maxwell- Boltzmann Distribution Function followed by ideal gas. Therefore Electrons at E ≥ (E_F + 0.26eV) behave like ideal gas molecules in a closed container. If we look closely we find that at energies equal or greater than (E_F + 0.26eV) ,

$n << N(E)dE$

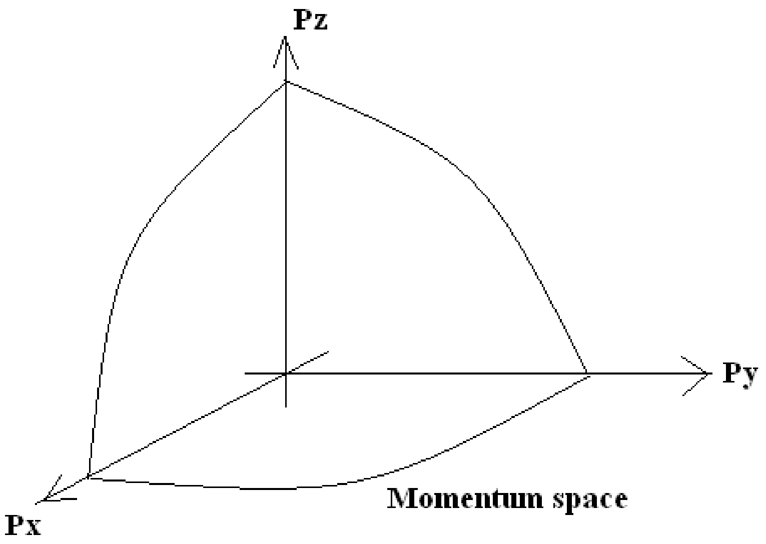
That is number of particles is much less than the available allowed states and this is known as NON-DEGENERATE STATES.

Therefore we can generalize that fermions in degenerate state follow Fermi-Dirac Statistics whereas fermions in non-degenerate states follow Maxwell- Boltzmann Statistics.

SSPD_Chapter 1_Part 17_Density of States

1.17. Calculation of Density of States in the Conduction Band of metals based on Quantum Mechanics.

Figure(1.88) describes the momentum space in terms of three reference momentum vectors p_x , p_y and p_z . Figure (1.89) describes the potential well created by one atom in (a), describes the overlapping potential wells resulting in periodic variation of potential along a linear array of atoms in (b) and a 3-D potential box in a single crystal solid with periodic potential variations in (c).



Figure(1.88) Momentum Space .

We saw in Chapter 1_Part 8continued_Electron in an infinite potential well that in a 1-D potential well(Eq.1.51):

$ka = n\pi$

where k is the wave vector or wave number, a is the lattice constant and n is the principal quantum number.

And $E = \hbar_2 k_2/(2m^*) = (h_2/4\pi_2)(1/(2m^*))(n\pi/a)_2 = n_2 h_2/(8m^* a_2)$

where m* is effective mass of electron.

Therefore $E = n_2 h_2/(8m^* a_2)$ **1.164**

In 3-D potential box where length , breadth and height are a, b and c

$E = (h_2/8m^*)[n_{x2}/a_2 + n_{y2}/b_2 + n_{z2}/c_2]$

If $a=b=c=L$ then $E= (h_2/(8m^* L_2))[n_{x2} + n_{y2} + n_{z2}]$

Or $E = (h_2\, n_2)/(8m^* L_2))$

Where $n_2 = n_{x2} + n_{y2} + n_{z2}$ **1.165**

At $n_x = 1$, $n_y = 1$, $n_z = 1$, $E_0 = (h_2/(8m^* L_2))(3)$

This is the lowest energy state and this is non-degenerate. Therefore the lowest energy state electron is non-degenerate.

But the energy state $E_1 = (h_2/(8m^* L_2))(6)$ has three fold degeneracy.

The following three states have the same energy namely:

$n_x = 2, n_y = 1, n_z = 1$

$n_x = 1, n_y = 2, n_z = 1$

$n_x = 1, n_y = 1, n_z = 2$

Therefore E_1 has three fold degeneracy.

Next energy state $E_2 = (h_2/(8m^*L_2))(1+4+9) = (h_2/(8m^*L_2))(14)$ has 6-fold degeneracy.

As the energy increases the degeneracy increases. From the allowed energy states and the number of electrons present we can determine if the given solid is conductor, semi-conductor or insulator.

1.17.1. Theoretical Formulation of density of permissible energy states.

Now we will do the theoretical formulation of density of permissible energy states.

Suppose

$N(E) =$ permissible energy states per unit electron-volt per unit volume;

Therefore $N(E)dE =$ permissible energy states per unit volume between E and $E + dE$.

We are studying a cubic volume with equal sides.

From Eq (1.165) we know that:

$$E = (h^2/(8m^*L^2))[n_x^2 + n_y^2 + n_z^2]$$

Here with positive integer values of wave numbers n_x , n_y, n_z we have quantum states associated. In a 3-D Cartesian Momentum Space with n_x , n_y, n_z chosen as frame of reference, in the positive octet as shown in Figure(1.90), every point with integer coordinates are permissible and possible quantum states.

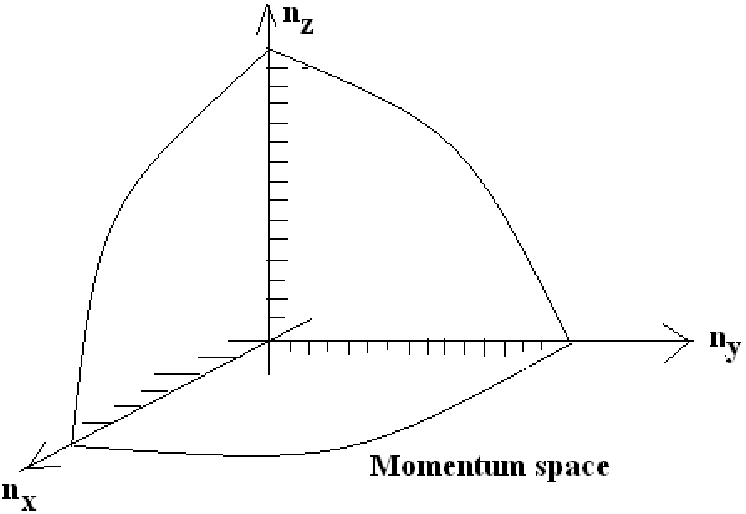


Figure (1.90) Positive Octet of 3-D momentum space with wave numbers n x , n y , n z as the frame of reference.

In reality n_x , n_y, n_z are discrete integer values but if we regard them to be continuous we will not incur much error. So we will regard n_x , n_y, n_z to be continuous.

Assuming that n_x , n_y, n_z are continuous we have the following relation:

$N(E)dE = (1/8)(4\pi n^2)dn$ =

permissible energy states in incremental energy dE in positive octet…………………………………………………… **1.166**

Where $4\pi n^2$ is the surface area of a sphere of radius n and $4\pi n^2 dn$ is the volume of the shell of radius n and thickness dn in n_x , n_y, n_z Cartesian Momentum Space. Since we are considering only the positive octet, we are considering only the positive values of n_x , n_y, n_z. Hence (1/8) factor has been introduced.

From Eq(1.165) $n^2 = (8m^*L^2/h^2)E$

Square root of this expression is:

$n = (8m^*L^2/h^2)^{1/2} E^{1/2}$…………………………………………………**1.167**

Differentiating n with respect to E we get:

$dn/dE = (8m^*L^2/h^2)^{1/2}(1/2)(1/E^{1/2})$ ………………………………**1.168**

Substituting the values of dn/dE and n^2 in Eq(1.166)

$N(E)dE= (1/8)(4\pi)((8m^*L^2/h^2)E) (8m^*L^2/h^2)^{1/2}(1/2)(1/E^{1/2})dE$

Simplifying the expression:

$N(E)dE= (\pi/4)((8m^*)^{3/2}/ h^3)(L^3)E^{1/2}dE$

Therefore permissible states per unit volume:

$N(E)dE=(4\sqrt{2})\pi((m^*)^{3/2}/ h^3)E^{1/2}dE$

Since every permissible state has two electrons of opposite spins therefore with every point having integer values of n_x , n_y, n_z has two quantum states associated. Therefore permissible quantum states per unit electron volt and per unit volume is:

$N(E)= (8\pi(\sqrt{(2m^*)})/ h^3)m$
$E^{1/2}$**.................................1.169**

In 3-D case $N(E) = CE^{1/2}$--------- as in case of bulk devices.

In 2-D case $N(E) = m^*/(\pi\hbar^2)=$ uniform---------- as in case of quantum sheet devices

In 1-D case $N(E)= C/E^{1/2}$------------------------- as in case of quantum dot devices.

Special Quantum Devices can be fabricated using the fact that Density of States variation law depends on the dimensions.

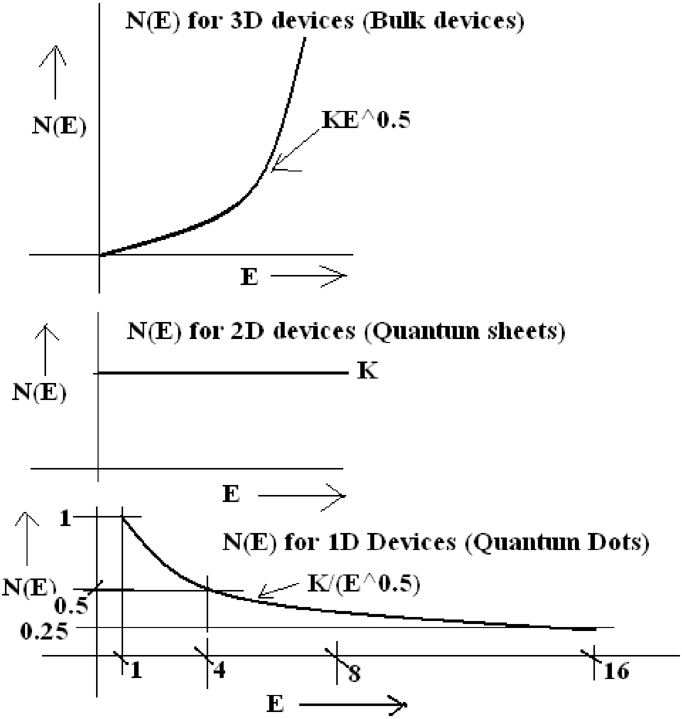


Figure 1.91. Density of States with respect to E in bulk devices, sheet devices and quantum dot devices.

1.17.2. DETERMINATION OF THE CONDUCTING ELECTRONS IN A METAL.

Let the number of electrons in conduction band be = n per cm^3;

Therefore $n = \int\{dn, 0, E_F\} = \int\{N(E)dE, 0, E_F\}$;

Substituting Eq.(1.169):

$n = \int\{(8\pi(\sqrt{(2m^*)})/ h^3)m^* E^{1/2}dE, 0, E_F\}$

$n = (8\pi(\sqrt{(2m^*)})/ h^3)m^* E_F^{3/2} \times (2/3)$……………………………………………… **1.170**

Rearranging the terms we get:

$E_F = (3n/8\pi)^{2/3}(h^2/2m^*)$ ………………………………………………………**1.171**

Table 1.33. The Fermi Energy (E F) and Work Function (W F) of selected Metals.

[From *Principles of Electronic Materials and Devices, Second Edition,* **S. O. Kasap(© McGraw- Hill, 2002)]**

Metal	Ag	Al	Au	Cs	Cu	Li	Mg	Na
W F (eV)	4.26	4.28	5.1	2.14	4.65	2.3	3.7	2.75
E F (eV)	5.5	11.7	5.5	1.58	7.0	4.7	7.1	3.2

SSPD_Chapter 1_Part 18_Kinetic Theory of Gases from QM perspective

SSPD_Chapter 1_Part 18 gives the Kinetic Theory of Gases from Quantum Mechanics perspective.

1.18. THE CORRECT INTERPRETATION OF KINETIC THEORY OF GASES BASED ON QUANTUM MECHANICS.

A thermally isolated system is known as an adiabatic system. In an adiabatic expansion we have the following law:

$PV_{\gamma} = \text{Constant}$
………………………………………………………………
1.172

$PV=kT$
…………………………………………………………………
1.173

$PV=(\gamma-1)U$ **where U(internal energy)=(3/2)kT**
…………………………………………………………………………
1.174

According to equipartition law of energy with every degree of freedom the average thermal energy associated is (1/2)kT.

A monoatomic molecule has only three translational degrees of freedom. Hence the average thermal energy associated:

$U = (3/2)kT$

Or $kT = (2U)/3$

Therefore in a monoatomic molecule gas from Eq.(1.173) and Eq.(1.174):

$(\gamma-1) = (2/3)$

or $\gamma = 1 + (2/3) = (5/3) = 1.667$

In a diatomic molecular gas, there are three degrees of freedom for translation, two degrees of freedom for spin around its longtitudinal axis and transverse axis and two degrees of freedom for vibration. Therefore total internal energy is:

$U = (3/2 + 2/2 + 2/2)kT = (7/2)kT$

Or $kT = (2/7)U$

Therefore in a diatomic gas again from Eq.(1.173) and Eq.(1.174):

$(\gamma-1) = (2/7)$

or $\gamma = (9/7) = 1.286$

Experimentally observed values of the adiabatic constant is tabulated in Table(1.34). The observed values of adiabatic constant cannot be explained on the basis of Classical Physics. Here again Quantum Physics comes to our rescue.

Table.1.34 Adiabatic Constants for different gas systems

Gas	T(°Celsius)	γ
He	180	1.660
Kr	19	1.68
H_2	100	1.404
Br_2	300	1.32
I_2	185	1.3
C_2H_6	15	1.22

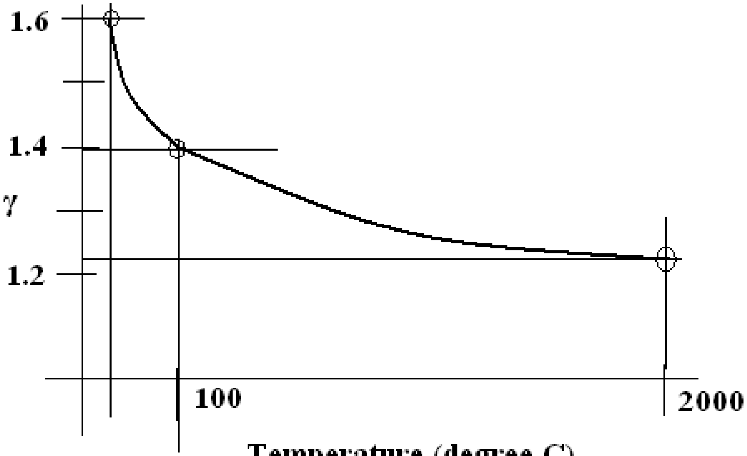


Figure 1.92 Variation of Adiabatic Constant with temperature for a given gas system.

According to Classical Physics many molecules can stay in the same energy state in the same volume. But according to Quantum Mechanics this is not permissible.

In a mono-atomic gas there are only three degrees of freedom of translation are available and these degrees of freedom is constant . Hence Adiabatic Constant is uniform with variation with temperature. This is found experimentally correct. But in diatomic molecules or multiatomic molecules this is not true.

Let us take diatomic molecule Hydrogen. At high temperature it has seven degrees of freedom. Two vibrational degrees of freedom, two degrees of freedom in rotational motion one is spin rotation and the other is orbital rotation and three degrees of freedom of translation . But as the temperature is reduced it loses the vibrational degrees of freedom and subsequently it loses rotational degrees of freedom. At low temperature only translational degrees of freedom remain. Therefore Adiabatic Constant is 1.6 at 25 degree Celsius. This falls to 1.4 at 100 degree Celsius. This further falls to 1.286 at 2000 degree Celsius.

The experimentally observed Adiabatic Constant variation with temperature as shown in Figure 1.92 is the most convincing vindication of the correctness of Quantum Physics at molecular scale.

Stefan- Boltzmann Law gives the total energy emitted by a Black-Body Radiator to be:

$E_T = \sigma T_4$……………………………………………………………………………
1.175

Classical Physics has no theoretical explanation for this Fourth Power Law. Quantum Mechanics gives the theoretical value of σ which is:

$\sigma = (2\pi_5 k_4)/(1.5c_2 h_3)$
…………………………………………………………………………
1.176

We saw Wien’s Displacement Law in Black Body

Radiation:

That is $\lambda_m = \Omega/T$ ……………………………………………………………… **1.177**

Classical Physics cannot arrive at the theoretical value of Ω. But Quantum Physics gives an elegant theoretical explanation of Black Body Radiation and arrives at the theoretical value of the constant : $\Omega = ch/(4.9651k)$

where $c = 3\times10^8$ m/sec = velocity of light in vacuum,

$h = 6.626\times10^{-34}$ J.sec = Planck's Constant,

$k = 1.38\times10^{-23}$ J/Kelvin;

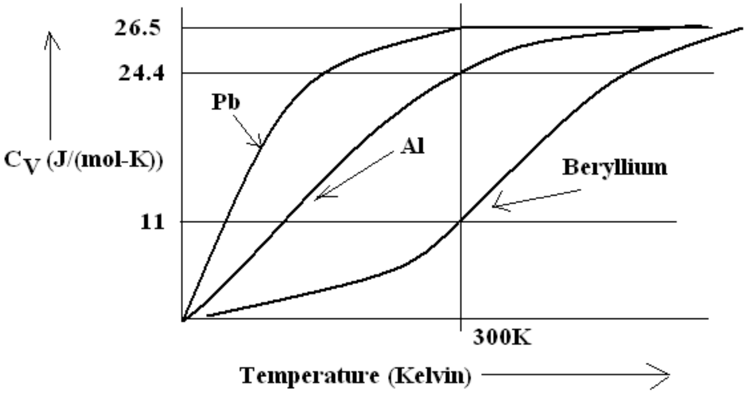
Substituting the Universal Constants in the above Equation:

The constant in Eq.(1.177) $\Omega = 3\times10^{-3}$m.Kelvin;

Rewriting Eq.(1.177) $\lambda_m = (3\times10^{-3}/T)$ m ……………………………………………… **1.178**

At 3Kelvin we observe 0.1cm microwave radiation pervading the whole Universe. This is called 3Kelvin Cosmic Microwave Background Radiation.

In Molar Specefic Heat theoretical formulation also, Classical Theory has no consistent model. Only Quantum Mechanics can give a consistent model.



Figure(1.93) Molar Specific Heat variation with temperature in Kelvin.

Table 1.35. Molar Specific Heat at 300 Kelvin for different monoatomic elements.

Elemental Solids	Molar Specefic Heat (at 300Kelvin)
Pb	26.5
W	24.8
Ag	25.5
Cu	24.5
Al	24.4

Beryllium	11	
Boron	13	
Carbon(Carbon)	6	

Average Kinetic Energy per degree of freedom = (1/2)kT;

Average Potential Energy per degree of freedom = (1/2)kT;

Since we have monoatomic elements therefore there are only three degrees of freedom of translational motion.

Therefore Molar Heat Q= N_A(3kT);

Therefore Molar Heat Capacity= dQ/dT = 3 N_Ak=3R

where R- Rydberg Constant and N_A= Avogadro’s Number;

Substituting the Universal Constants:

Molar Specific Heat of a given material =

Molar Heat Capacity of the given material/Molar Heat Capacity of water=24.93 J/(Mol.K) ;

Therefore all Monoatomic elements should have Molar Specific Heat= 24.93J/(Mol.K).

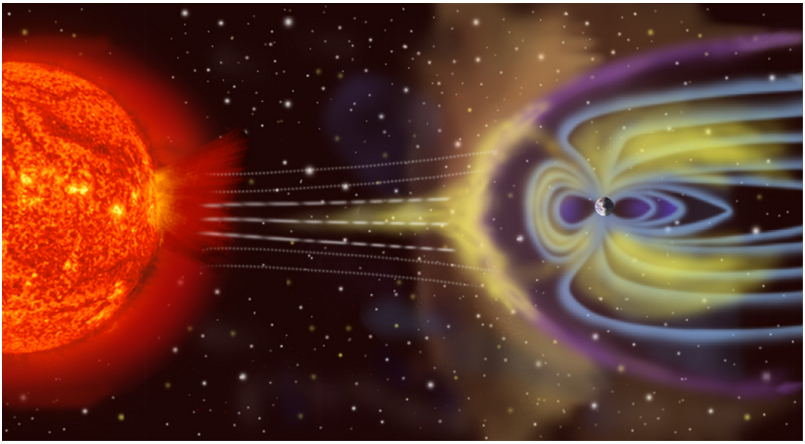
SSPD_Chapter 1_Appendices_Space Weather Research
This is one of the Appendices of SSPD_chapter 1.
This gives the physics of Space Weather Research.

SUN-SOLAR ACTIVITY & COMMUNICATION

Between the ground and interplanetary space:

1. The troposphere- human population is limited to this level and uses only the first 6 km of atmosphere, a thin layer that has to be reserved.
2. The stratosphere-(between 15km to 45km) very dry, it is characterized chiefly by ozone layer which is an effective barrier to ultra violet light from sun.
3. The mesosphere-(between 45km to 80 km) An icy layer(-90 degree centigrade at the top of this layer)
4. The Thermosphere-(between 80km to 480km).very hot layer(from 850 to 2300 degree centigrade depending on solar activity). This contains a layer of plasma called ionosphere. Ionosphere acts as an effective mirror for Radio Waves-Short Wave i.e. 1MHZ to 30 Mhz.

The Earth behaves like a huge bar magnet with the magnet almost but not exactly aligned along the spin axis. This bar magnet creates a Magnetic Field terminating at the north pole and originating at the South Pole. From a height of 1000km an immense magnetic field surrounds the Earth as shown in the Figure below. This 3-d magnetic field which under normal condition would have a cylindrical symmetry but in presence of solar wind it is flattened on the windward side and elongated on the leeward side as shown. This is called MAGNETOSPHERE. Earth is continuously bombarded by fierce solar Wind and Cosmic rays. This magnetosphere repels the stars high-energy radiation.



Figure(1)The Gaint Magnetic Field of Earth Bar Magnet –Magnetosphere.

SUN-A gigantic Nuclear Fusion Reactor.

Our Sun is a humble star in the Milky Way Galaxy of stars. it was born about 4.56 Billion years ago and it will continue to react for another 5 billion years. That is its total life span is 10 billion years.

Once its nuclear fuel runs out, it will expand to Red Giantand then collapse into White Dwarf. The solar System will be wiped out.

SUN- Its Present Scenario.

The fusion reaction between hydrogen and helium is the primary source of energy. Sun is a main sequence star where a balance is maintained between the gravitational inward collapse and outward expansion due to radiation pressure of the thermonuclear energy released due to Hydrogen atoms fusion. This state of equilibrium has been maintained since the beginning and it will be maintained for the whole life span of 10 billion years.

THE BURNING CORE

At the core of the sun, temperature is 15 million degree Celsius and a pressure equivalent to 160 times the atmospheric pressure.

In the heart of the Sun , 500 million tones of hydrogen fuses into helium Gas . This fusion reaction leads to Mass defect of 4 million tones which by $E=mc2$ relationship is converted into energy. This released energy radiates to the adjoining zone. From there through convective current of gases it travels to the surface of the sun.

SOLAR PHOTOSPHERE

Beyond the convective layer is solar photosphere. This layer is 200 km thick on an average. Here the matter is cool enough to turn from plasma state to neutral state. Therefore radiation decouples from matter and freely travels out.

The photosphere has a mosaic of bright spots called granulations plus sporadically appearing darker patches called SUNSPOTS. At the granulations areas the photosphere is 1000 to 2000 km thick.

The average temperature of the photosphere is 6400 degree centigrade but at the Sun-spot it is cooler at 2,000 to 3,000 degree. Here magnetic fields are emanating which are 25,000 times stronger than earth magnetic field. Sun Spots appear and disappear. They are surrounded by clouds of bright, burning gas called faculae. Sun Spots number and sizes are indicative of Sun's Solar Activity. Larger sizes and more numerous number is indicative of Sun's Activity. The Sizes and number wane and wax in cycles of 11 years. This is known as 11 years cycle of Sun Activity. By studying the sun spot, the rate of rotation of Sun has been determined to be 27 days.

SOLAR CORONA- SOLAR FLARES-CORONAL MASS EJECTIONS

There is a pink fringe several thousand kilometers broad, around the solar disc. This Pink fringe is called the chromospheres. As the solar activity increases, magnificent prominences shoot out from this layer. They look like curls or bushes. These are Coronal Mass Ejections (CME). In these coronal mass ejections the sun blasts billion of tons of plasma blob material at very high speed. These prominences are sculpted by the magnetic field. The gas is ejected and re-circulated back at a high speed of 1000 km/sec. Sometimes this velocity is as high as 500,000km/sec.

Sudden explosions of Solar Atmospheres release energy equivalent to billions of Atomic Bombs within few minutes. The abrupt rearrangement of solar magnetic field , which arch upward from the surface of the sun, lead to this sudden release of energy.CME are accompanied with burst of X-Rays. Solar flares and different aspects of Space Weather involve the interplay of magnetic field and plasma.

The exact physics has been unraveled by the study of Yohkoh satellites & Ramaty High Energy Solar Spectroscopic Imager(RHESSI). Solar flares erupt from Sun Spots. At Sun Spots the solar magnetic fields is much stronger. It is the magnetostatic energy of these powerful magnetic fields which are unleashed during solar flares as thermal energy. Oppositely directed magnetic field lines come together and partially annihilate and rearrange. This is known as RECONNECT. This magnetic reconnection is responsible for the solar flare and CME. Invisible Magnetic Loops are emanating and terminating at the sun spots. These magnetic loops trap hot plasma. Initially this plasma is at several million Kelvin temperature. Hot charged particles spiral around the magnetic lines of force to and fro between the two foot points. The acceleration and deaccleration lead to synchrotron radiation in form of extreme UV and hard X-ray. In the process plasma heats up to 10 to 40 million Kelvin. As the magnetic reconnect takes place, gigantic solar flares erupt and CME occurs. As the sun spot increases , number of magnetic loops increase and hence the magnetic reconnect becomes more numerous leading gigantic solar storms.

Beyond the chromospheres, corona extends up to 10 Solar radii. This has long jets of gas running through it. A large number of energetic ionized hydrogen particles manage to penetrate the magnetosphere. These particles get trapped in the magnetosphere and travel in toroidal path around the magnetic field lines. This belt of trapped particles is called VAN ALLEN BELT.

These ionized particles keep travelling between North & South Magnetic poles of the Earth. At a height from 100 km to 1000 km the inward – rushing or outward –rushing charged particles interact with neutral atmospheric gas particles(Nitrogen & Oxygen).In the process they produce the peak of solar activity. Auroral displays become brighter and solar perturbations are perceptible up to the surface of the Earth.

Depression in corona let matter escape at a speed of 800km/sec. The brief solar flares cause a wind of 1000km/sec. Whereas the eruptions of prominences cause the solar wind to speed up to 2000km/sec.

11 year cycle remains an enigma.

EFFECT OF SOLAR ACTIVITY ON POWER TRANSMISSION & COMMUNICATION

Periodic increase in solar activities in form of increased sun-spots and furious sun-flares lead to major disruptions in power transmissions and communication networks.

August 2, 1972 a 230KV transformer exploded in British Columbia in Canada.

On March 13,1989 , at the peak of Solar Activity, Quebec Province went without power due to failure of the electric transmission network.

October-November 2003 witnessed some of the largest solar flares ever recorded in mankind's history. The barrage of charged particles forced many satellites to be turned off. Despite this preventive measure many satellites got damaged. Astronauts in international Space Station had to retreat to better shielded portions of their service modules. The air-liners were routed to lower latitudes to prevent communication failure and to protect passenger from unsafe level of radiation exposure. National Transmission Grids had to properly protected against surges. Despite this 50,000 residents in south Sweden faced power failure briefly.

The Global Radio Broadcast at Short Wave RF gets seriously effected due to disruptions of ionosphere surrounding the Earth.

Manned Space Missions can get disrupted. Understanding the solar flares will help the space walkers and astronauts going to Moon &Mars in the coming decade. Astronauts exposed to eruptions can receive a year's worth of radiation.

Coronal mass Ejections cause Solar Storm. This distorts Earth's magnetic field. Backlash of charged particles trapped in radiation belts generates massive voltage that can damage satellites and power grids.

THE PATTERN OF SOLAR ACTIVITY

(cooling Sun may bring relief to sweltering Earth-The Hindu)

Sun's output varies over a cycle of 11 years. When vast amount of energy are being released deep inside & strong magnetic fields rip through the sun's surface, this is the period of heightened solar activity accompanied with increased number of sun-spots. From 1650 to the end of 1700, the sun-spots almost completely disappeared & Earth experienced little ice again. At this time Thames froze over, Europe shivered & harvest failed.

In the next few years from now, solar activity will come down . This will lead to increase in cloud formation & reduction in UV radiation reaching the atmosphere. Overall it will bring about a decline of 0.2 degree centigrade in Global Temperatures.

It is predicted the Carbon Emissions of 450ppm or Carbon Equivalent Emissions of 500ppm will bring a 2 degree C increase accompanied with great floods.

Carbon Dioxide is not the only greenhouse gas. Methane, Nitrous Oxide & Hydrofluorocarbons are the other greenhouse gases. When we include the effect of other green house gases then we talk in terms of Carbon Dioxide Equivalent.

STEREO (Solar Terrestrial Relations Observatory) mission to scan the SUN In 3-D:

NASA is to launch two nearly identical space-based observatories that will enable the construction of 3D view of the Sun. This is a $550-million, two year mission. This will help us understand CORONAL MASS EJECTIONS better.

NASA Scientist Madhulika Guhathakurta says,” of the bazillion stars that we have in our night sky, sun is the one that only counts. Any understanding or breakthrough we can make in understanding the sun & its environment is of direct relevance to every human being on this Planet”.

This mission will improve early warning system against CME s.

[“The Mysterious Origins of Solar Flares” George D. Holman, Scientific American, April 2006;

“The fury of Space Storm”, James L. Burch, Scientific American, April 2001;

”Paradox of Hot Sun’s Corona” Bhola N. Dwivedi & Kenneth J.Phillip, Scientific American, June 2001”.

http://hesperia.gsfc.nasa.gov/sftheory/]

Crystal Growth-Bulk and Epitaxial Film-Part 1
Crystal Growth Part 1 goes on to describe the preparation of electronic grade polycrystalline Si from sand and then preparing single crystal pure or doped Silicon Ingot with a given crystalline orientation.

CRYSTAL GROWTH

- (1)Preparation of polycrystalline Electronic Grade Silicon material in *Siemens Reactor*.
- (2)Bulk Crystal preparation or Pure,Electronic Grade, Single Crystal Silicon ingot preparation by *Czochralski method* and by *Float Zone Method*.
- (3)Diamond cutting of Silicon Ingot into Silicon Wafers and *lapping, polishing and chemical etching* of the Wafers to obtain mirror finished Substrates for IC preparation.
- (4)Epitaxial Film Growth-*Chemical Vapour Phase Deposition*(CVD) or *Liquid phase Epitaxy*(LPE).
- (5)For Photonic Application , compound Semiconductor Epitaxial Film Growth is achieved by *Molecular Beam Epitaxy* (MBE).

PREPARATION OF POLYCRYSTALLINE ELECTRONIC GRADE SILICON MATERIAL.

Quartz Sand (SiO_2) plus Coke are put in a container with submerged electrode **Arc Furnace**. Electric energy is consumed at the rate of 13 kWhr per kg and SiO2 is reduced to Silicon by Coke. This silicon is of 98% purity. The solid silicon is pulverized and kept in an oven where Hydrogen Chloride vapour is passed at 300ºC. Silicon is converted into SiHCl3 (liquid) which is called Trichloro Silane.

SiO_2(Quartz Sand) + 2C (coke) = (at high Temperature)Si + CO_2

Si(pulverized form) + 3HCl (vapour) =(At 300ºC) $SiHCl_3$ (liquid) + H_2↑

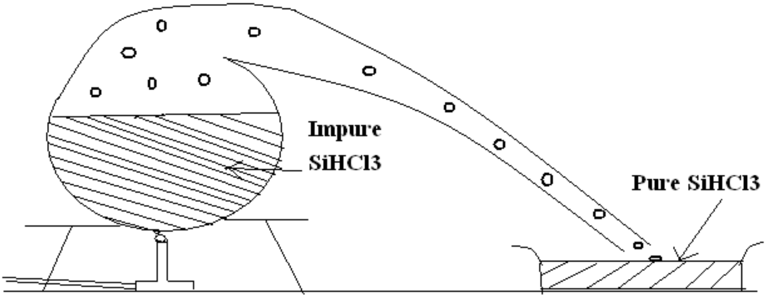


Figure 1. Multiple Distillation Set-up for preparing electronic grade TriChloroSilane.

$SiHCl_3$(liquid) has B.P 31.8ºC. Whereas most impurities are less volatile. Therefore by multiple distillation electronic grade TriChloroSilane is obtained which eventually is used in SIEMEN'S REACTOR.

H_2 + $SiHCL_3$ =(at 1000ºC) Si + 3HCL

(Hydrogen reduction)

(a form of CVD of Silicon on slim rod of Silicon on Tantulum wire)

Slim rod contains Tantalum Wire surrounded by Silicon as shown in Figure 2. Slim rod can be formed either by CVD on hot Tantalum wires or can be pulled from the melt.

(When Si deposited by CVD on hot Tantalum wires then there can be thermal run away due to negative temperature coefficient of resistance(t.c.r) of surrounding silicon deposit dominating over positive t.c.r of the Tentalum wires. The thermal runaway has to be prevented by controlling the current.)

The resulting fattened slim rod has considerable metal contamination at the core of the rod.This can be removed by dissolving the central metallic core with acid. ANitric/Hydrofluoric acid mixture is then pumped through the hollow core. This cleaning method widens the hollow core by removing silicon and metal contamination(1mm/hr). First nitric acid is used to dissolve the metallic core . Next Hydrofluoric acid is passed through the hollowed core to dissolve the inner layer of silicon and in the process removing the metal contamination. The fattened, purified rod of silicon thus obtained will be used as the feedstock for crystal pulling. Hydrofluoric acid is very corrosive and it can effect our bones even . Therefore Teflon gloves , Teflon Aprons,Teflon tweezers and Teflon beakers have tobe used while working with HF acid.

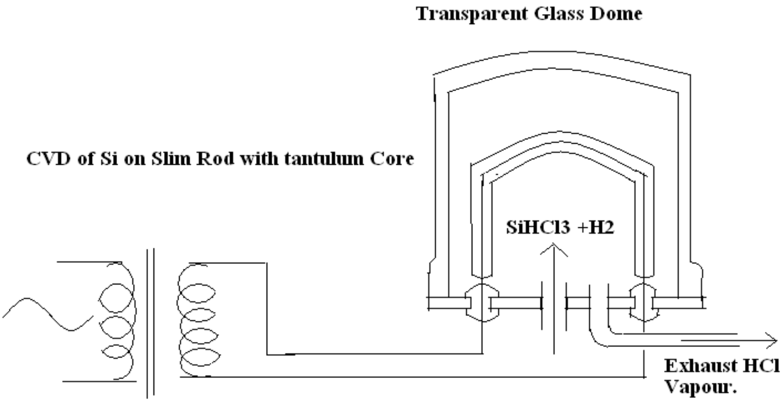


Figure 2. Siemen's Reactor for Chemical Vapour Deposition of Electronic Grade Si on Tantulum Wire Core which is heated at 1000 degree centigrade.

FLOW CHART OF SIEMEN’S REACTOR

Reduction of sand with carbon gives impure polycrystalline Silicon

↓

Reaction of pulverized raw silicon with HCl gaseous vapour to form TriChloroSilane

↓

Multiple distillation of TriChloroSilane to obtain purified electronic grade TriChloroSilane

↓

Thermal decomposition of SiHCl3 at 1000 degree centigrade in Sieman’s reactor to obtain fattened rods of electronic grade Silicon

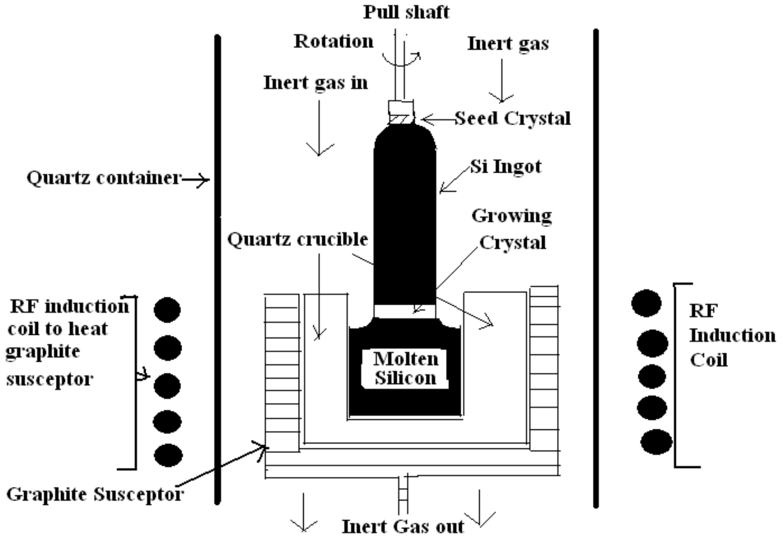


Figure 3. The schematic illustration of the growth of a single crystal Si Ingot by Czochralski Technique.

Figure 3. The schematic illustration of the growth of a single crystal Si ingot by the Czochralski technique.

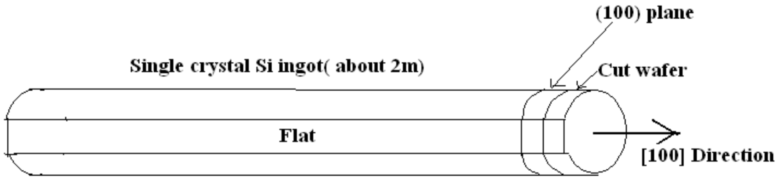


Figure 4.The crystallographic orientation of the Silicon Ingot is marked by grounding a Flat. The ingot can be as long as 2 meter. Wafers are cut using a roating annular diamond saw. Typical wafer thicknes is 0.6-0.7 mm.

The crystallographic orientation of silicon ingot is marked by grounding a flat as shown in Figure 4. The ingot can be as long as 2m. Wafers are cut using a rotating annular diamond saw. Typical wafer thickness is 0.6-0.7mm

CZOCHRALSKI METHOD OF SINGLE CRYSTAL PULLING

Polycrystalline silicon is not suitable for Electronic Device preparation. Polycrystals have grain boundaries which make mobility unpredictable because of unpredictable amount of defect scattering. Grains boundaries cause uneven distribution of dopents. Along the grain boundaries there is more rapid diffusion of dopent as compared to that in the bulk. Hence we opt for single crystal silicon

Czochralski Equipment, as shown in Figure 3 and Figure 5, has a graphite lining. Graphite is of Nuclear Reactor Grade. Within the Graphite lining, Quartz crucible of high purity is placed for holding the silicon melt. Induction heating is used for maintaining the silicon melt at 1420ºC . Induction heating is eddy current heating. The eddy currents are induced in the graphite lining by alternating magnetic field caused by RF induction coil. Crystal pulling is carried out in inert atmosphere of argon to prevent oxidation and to suppress evaporation.

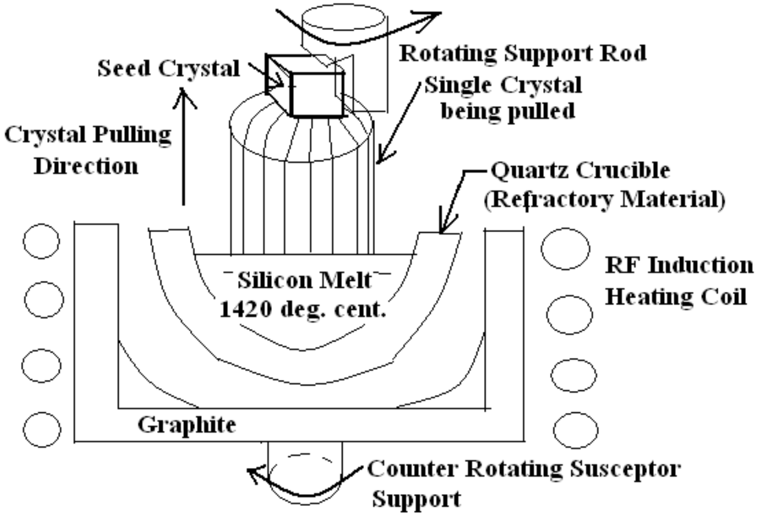

Rotating Support Rod
Seed Crystal
Single Crystal being pulled
Crystal Pulling Direction
Quartz Crucible (Refractory Material)
RF Induction Heating Coil
Silicon Melt 1420 deg. cent.
Graphite
Counter Rotating Susceptor Support

Figure 5. Czochraski Equipment.

Seed crystal determines the perfection and the orientation of the single crystal being pulled. Hence seed is dislocation free and of desired crystal orientation.

The total length of Si ingot is L=2m & Φ =15cm, 20cm, 25cm, Optimum pulling rate is 2 mm per min. In 24 hr period the total length of ingot is pulled .Silicon being highly refractive (m.p.1410 c) is highly contaminable. The Quartz of the containment vessel undergoes dissolution into Si during its growth. The result is the inclusion of O2 as a donor impurity into Silicon. Carbon is another impurity that finds its way into Silicon. These contaminants are particularly harmful for powerful devices.

Counter rotation and rotation are produced to provide uniformity , homogeneity of thermal effects and dopents. This minimizes the defects also.

Inspite of these homogenizing effects there are swirl effect which lead to non-uniform distribution of desired and undesired contaminants. This leads to non-uniform resistivity distribution across a silicon wafer. The spatial variation is such that these non-uniformities are not particularly important in I.C. or in miniature devices. However in power devices ,which are large area devices , local variation in doping can lead to non-uniform heating and hence to hot spots and failures. For example 3” diameter Thyristor carries 3000 amps. Even one defect per slice will be catastrophic to such a device but is inconsequential to I.C. Fabrication.

This implies that for applications in which contaminants are critical, crystal, wafers produced by CZOCHRALSKI METHOD are not suitable.

One way to avoid the contaminants is to be avoid contact with quartz crucible. This is achieved by FLOAT ZONE METHOD of Crystal Growth.

FLOAT ZONE METHOD OF CRYSTAL GROWTH

An electronic grade , pure polycrystalline silicon rod is held vertically as shown in Figure 6 in inert atmosphere .thus no crucible is used and hence contaminations are avoided.

A seed crystal is held at the bottom of the rod to determine the crystallographic orientation

Through R.F induction heating a narrow zone of vertical, purified, poly crystalline bar is melted. Dimensions of the molten zone are such that surface tension hold it in place. Because of less than unity segregation coefficient , a(m)=(concentration M in solid phase)/(concentration of m in liquid phase), impurities prefer to stay in liquid phase.

As the R.F. coil is sifted from bottom to top , the impurities of the region are swept to the top and the recrystallized portion is relatively pure and single crystalline with orientation that of seed crystal. If high purity is required then several passes can be made and ends containing impurities can be sliced off leaving behind very pure single crystal silicon ingot. Pure crystal of very high resistivity can be achieved by repeated recrystallization through several passes.

Table 1. Segregation Coefficient for different dopents in different hosts.

Semiconductor	N Type	P type
Silicon	a(P) = 0.35,	a(B) = 0.8

Gallium Arsenide	a(As) = 0.3 a(Se) = 0.10	a(Zn) = 0.42

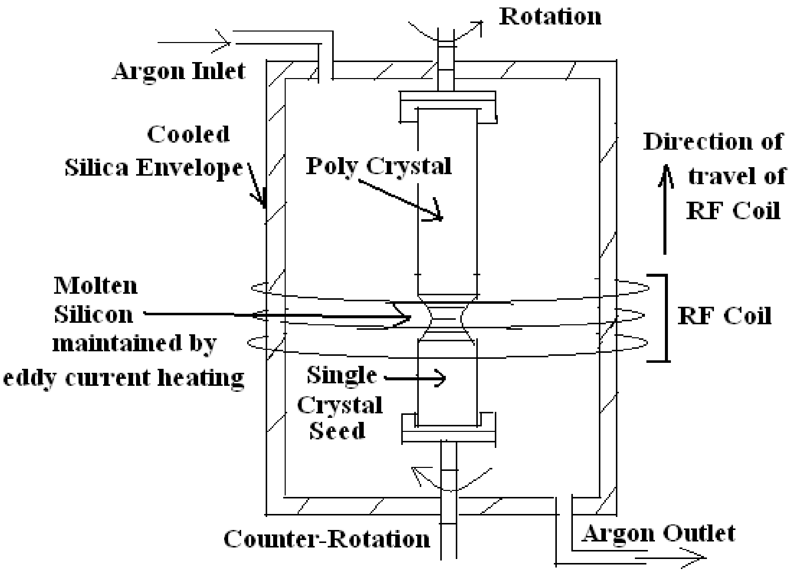


Figure 6. Float Zone Method of Crystal Growth.

Crystal Growth-Bulk and Epitaxial Film-Part 2
Part 2 of Crystal Growth describes the sawing of the Ingot into taper free, flat Si Wafers. Lapping , chemical etching and polishing are described to obtain mirror finish, defect free 300mm diameter Silicon Wafer.

Crystal Growth- Bulk and Epitaxial Film- Part2.

SLICING OF SINGLE CRYSTAL INGOT INTO SILICON WAFERS

Slicing of the Si ingot into Si wafers is achieved by circular saw blade. The circular saw blade is illustrated in Figure 7.

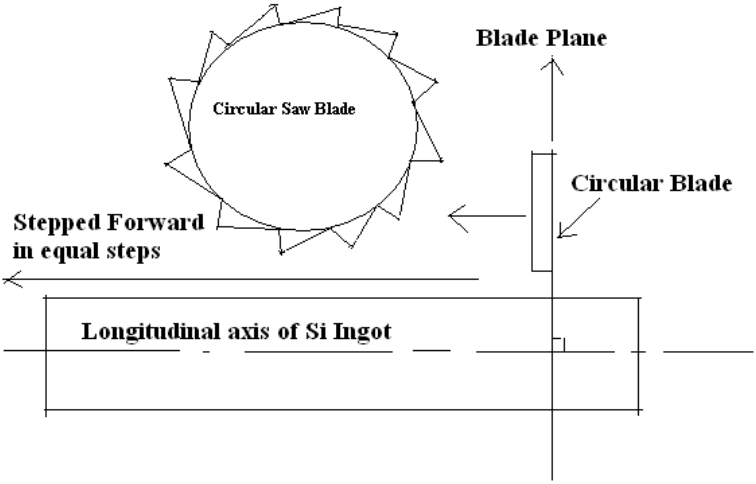
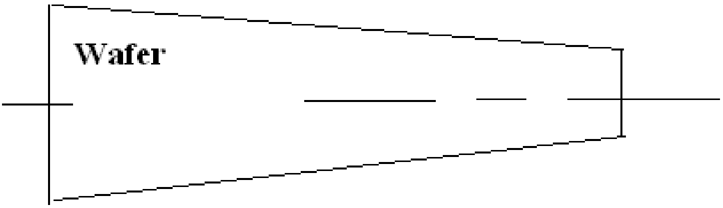
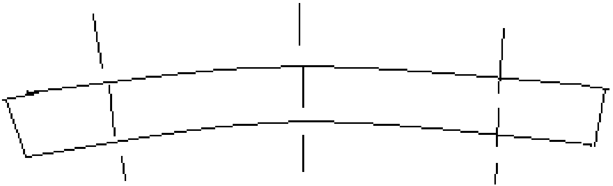


Figure 7. Circular Saw Blade with diamond impregnation. Saw Blade must be kept perpendicular to the lontitudinal axis of the ingot to avoid taper in the wafer.

Circular saw blade consists of stainless steel impregnated with diamond dust at the cutting edge. On the hardness scale diamond is the hardest. 500 micrometres thick Si wafers are sliced out through damage free and parallel sawing. The ingot should be sawed in parallel planes to avoid taper of the wafer.



Non-uniform thickness of the wafer gives rise to TAPER defect



By BOW or CURVATURE of Wafer , flatness is effected.

Figure 8. TAPER & FLATNESS of Wafer

In the process of sawing , mechanical, defects are inevitable. To remove the mechanical defects and to achieve taper-free, flat, mirror finish wafers we do lapping, chemical etching and polishing.

LAPPING: For removing mechanical scratches and abrasions encountered during sawing, lapping is essential. A batch of wafers are placed between two parallel plain steel discs which describe planetry motion in relation to each other. A slurry of cutting compound (generally alumina) is used and as lapping progresses, the grade of abrasion is refined. Smoothness is within 10 μm and flatness is within 2 μm.

CHEMICAL ETICHING: To remove edge damages which had been caused due to sawing and grinding, the wafers is dipped in HF acid using Teflon beakers. Chemical Vapor etching can also be done as described in epitaxy section.

POLISHING: A slurry of silica in NaOH is used for polishing to mirror smoothness. Lapping , Chemical Etching and Polishing removes 100μm thick substrate leaving behind 400μm thick silicon substrate on which I.C. Fabrication can be carried out..

FLOW CHART OF WAFER PREPARATION

Starting Material - Sand, Sand reduced to Silicon, Silicon converted to TriChloroSilane, through multiple distillation TrichloroSilane is purified to Electronic Grade and reduced in Siemen’s Reactor to Pure Electronic Grade Polycrystalline Si.

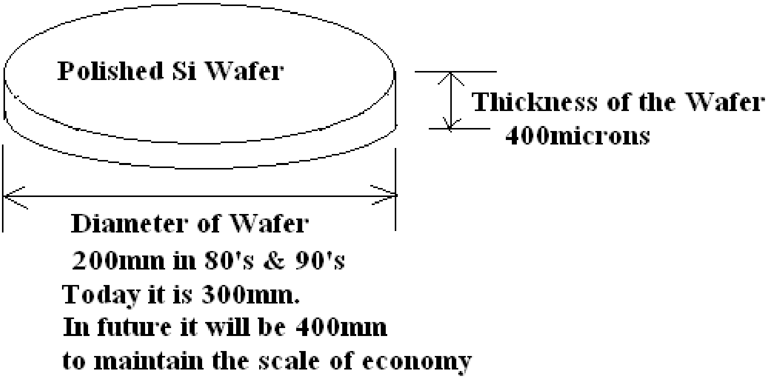
↓

From purified Silicon Melt, single crystal ingot is pulled out by Czochralski Method or Poly Si ingot is purified re- crystallized by Float Zone Method. Required Dopent (Phosphoprous for N Type ingot and Boron for P Type ingot) is added to the melt in calculated manner to form single crystal ingot of specific impurity type doping and of given resistivity.

↓

Sawing, Lapping, Grinding and Polishing

↓



In early days when IC Technology had just started i.e. in the year 1961 the wafer size was only 1 inch. Then it was: * 2 inch (50.8 mm). Thickness 275 µm. * 3 inch (76.2 mm). Thickness 375 µm. * 4 inch (100 mm). Thickness 525 µm. * 5 inch (127 mm) or 125 mm (4.9 inch). Thickness 625 µm. * 150 mm

(5.9 inch, usually referred to as "6 inch"). Thickness 675 µm. * 200 mm (7.9 inch, usually referred to as "8 inch"). Thickness 725 µm. * 300 mm (11.8 inch, usually referred to as "12 inch" or "Pizza size" wafer). Thickness 775 µm. * 450 mm ("18 inch"). Thickness 925 µm (expected). By all these technique , high purity (99.999%) single crystal ingots with a controlled amount of N or P type dopent can be produced. The maximum resistivity obtained in Si ingot is 10^4 Ω-cm whereas in Ga As and InP the $\rho max = 10^8 \Omega$-cm. That is almost semi-insulating ingot of GaAs and InP can be obtained. Semi-insulating GaAS and InP wafers are suitable for reduction of parasitics, for isolaton of active devices from one another, for reduction of power drain and for achieving high speed.

Presently because of technical constraints we have the option of Si, GaAs and InP substrate only. Therefore we have a very limited choice of overlay films also. This severely constrains the photonic device design.

Crystal Growth-Bulk and Epitaxial Film-Part 3

Part 3 of crystal growth sums up the properties of Silicon, describes the identity marks of Silicon Wafers in terms of Semiconductor Type and in terms of Crystal Orientation and this part defines the crystal orientation and its relevance in IC fabrication.

CRYSTAL GROWTH

Silicon, (Si):The most common semiconductor, atomic number 14,

energy gap Eg = 1.12 eV- indirect bandgap;

crystal structure- diamond, lattice constant 0.543 nm,

atomic concentration 5 x 10^{22} atoms/cm^{-3},

index of refraction 3.42, density 2.33 g/cm^{3}, dielectric constant 11.7,

intrinsic carrier concentration 1.02 x 10^{10} cm^{-3},

bulk mobility of electrons and holes at 300°K: 1450 and 500 cm^{2}/V-s,

thermal conductivity 1.31 W/cm°C,

thermal expansion coefficient 2.6 x 10^{-6}°C^{-1}, melting point 1414°C; excellent mechanical properties (MEMS applications);

single crystal Si can be processed into wafers up to 300mm in diameter.

In future this diameter will be 450mm.

P type= Always Boron (B) Doped N type= Dopant typically as follows:Res: .001-.005 Arsenic (As)Res: .005-.025 Antimony (Sb)Res: >.1 Phosphorous (P)

EPITAXIAL CRYSTAL GROWTH

The substrate or the wafer only constitutes the strong base of the integrated circuit. The actual active and passive components fabrication and there integration are carried out in overlay films which are grown by epitaxial technique.

EPITAXY is a Greek word meaning : ‘epi’ (upon) & ‘taxy’ (ordered). That is an epitaxial film, a few µm thick, is an orderly continuation of the substrate crystal. It grows very slowly layer by layer. Hence the dimension , defects and doping magnitude as well as uniformity can be precisely and accurately controlled in the crystal growth direction.

This precise control is obtained in Molecular Beam Epitaxy (MBE) but not in Liquid Phase Epitaxy(LPE) or in Chemical Vapour Phase Epitaxy (CVPE). The thickness accuracy is within ±3À which is essential for growing Quantum Photonic Devices namely Quantum Dots, Quantum Wells and Super-lattices

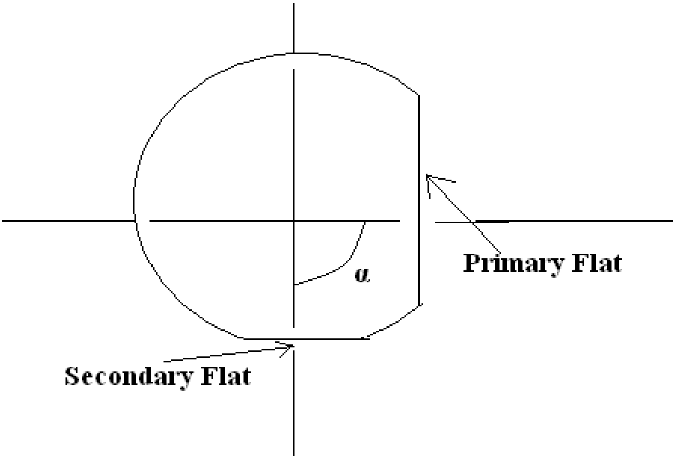


Figure 9 Silicon Wafer Identity Marks.

Table 2. Identity marks of the Wafer to identify its orientation and semiconductor type.

α (angle between primary and secondary flats as indicated in Figure 9)	Type	Orientation
45º	N	<111>
90º	P	<100>

180º	N		<100>	
0º	P		<111>	

The normal to the plane along which crystals cleaves is the cleavage plane orientation. Suppose the cleavage plane orientation is <111>. Miller Index is being used to define the planes and their normal. Figure 10 illustrates the Plane’s Miller Index and how the normal to the plane is represented. If the exposed surface of the Si wafer, which is known as major flat, is parallel to cleavage plane then the given wafer has a crystal orientation <111>.

If the cleavage plane orientation is <100> and the wafer major flat is parallel to YZ plane then the crystal orientation is <100>. In this case cleavage plane lies in YZ plane i.e. [100] plane and its orientation is perpendicular to YZ plane i.e. x-axis. Hence Wafer Crystal Plane orientation is <100>

Scribing the wafer along cleavage planes allows it to be easily diced into individual chips (‘die”) so that billions of individual circuits or systems on an average wafer can be separated into individual dies. Each individual die is eutectic ally bonded on ceramic substrate. The substrate is bonded to the header.The gold wire is connected to the bonding pads of the die on one end and to the chip terminals on the header by Thermo-compression bonding or by Ultra-sonic bonding. Next the die is hermetically sealed into Dual-in-Line(DIP) package or TO5 package

In <100> crystal orientation, scribed pieces form rectangle whereas in <111> crystal orientation, scribed pieces form triangles. Here we have to scribe from the base of the triangle to the apex.

For MOS fabrications, wafers with crystal orientation <100> are used. This helps achieve a lower threshold voltage. For BJT and other applications wafers with orientation <111> are preferred.

Silicon Crystal Bulk is isotropic to diffusion of dopents and to etchents used for etching the oxide layer. This is because of the symmetric property of Cubic Structure of Si. But real devices are built near the surface hence the orientation of the crystal does matter.

In 111 crystal terminates on 111 plane and in 100 it terminates on 100 plane. 111 plane has largest number of Si atoms per cm2 whereas 100 has the least number of atoms per cm2. Because of this difference 111 planes oxidize much faster because the oxidation rate is proportional to the Silicon atoms available for reaction.

But because the atom surface density is the highest the dangling bond surface density is also the highest in 111 hence Si/SiO2 has superior electrical properties in terms of interface states in 100.

Interface states give rise to 1/f noise or flicker noise. Because of this superiority all MOS devices use 100 crystal orientation. But historically BJT have used 111 because 111 crystal growth is easier to grow by Czochralski method. But as we move to sub-micron and deep and ultra-deep sub-micron BJT, 100 crystal orientation seems to be the crystal orientation of choice for BJT also.

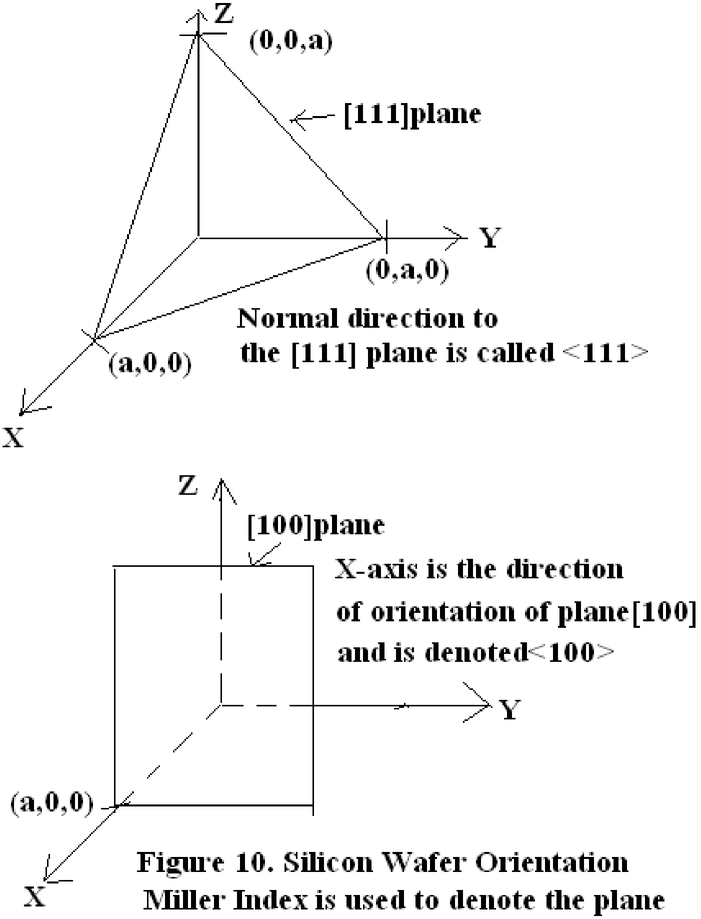


Figure 10. Silicon Wafer Orientation
Miller Index is used to denote the plane

Crystal Growth-Bulk and Epitaxial Film-Part 4
Part 4 of Crystal Growth describes the three technologies of Epitaxial Film Growth- Chemical Vapour Phase Deposition, Liquid Phase Epitaxy and Molecular Beam Epitaxy. First two are for Si Epitaxial Film growth and MBE is for compound semiconductor epitaxial films.

Crystal Growth-Bulk and Epitaxial Film_Part 4

EPITAXIAL FILM GROWTH.

There are two major epitaxial film growth technologies:

Silicon Epitaxial Film Growth Technology (Low End Technology hence economical) and Compound Semiconductor Epitaxial Film Growth Technology (High End Technology hence expensive and suitable for niche applications).

Silicon Epitaxial Film Growth Technology are further divided into:

Chemical Vapour Phase Deposition(CVPD) Technology and Liquid Phase Epitaxy(LPE). LPE is relatively cheaper but very toxic.

Compound Semiconductor Epitaxial Film Growth is achieved by Molecular Beam Epitaxy System which is inordinately expensive and in India in a few places only we have MBE systems namely TIFR(Mumbai), Solid State Physics Laboratory(Delhi), Central Electronics Engineering Research Institute (Pilani), IIT(Madras) and CSIO(Chandigarh).

In Table 3 we make a comparative study of LPE, CVPD and MBE. Molecular Beam Epitaxy is carried out under Ultra High Vacuum Conditions. This means we are working at 10^{-10} Torr where 1 Torr is 1mm of Hg. For achieving Ultra High Vacuum we have to work in three stages: Rotary Pump is used for achieving 10^{-3}Torr. Silicon Oil Vapour Pump is used to achieve 10^{-5}Torr and Ion Pump is used to achieve Ultra High Vacuum of 10^{-10}Torr. This ultra high vacuum requirement makes MBE equipment inordinately expensive. This equipment is imperative for Photonic Devices.

CHEMICAL VAPOUR PHASE DEPOSITION(CVPD) OF EPITAXIAL FILMS or Chemical Vapour Phase Epitaxy(CVPE).

In Figure 11 , we describe CVPD system for obtaining Si Epitaxial Films. A controlled chemical vapour of precise chemical composition at a precisely controlled flow rate is passed over silicon substrates. Si Substrates act as the seed crystals. They are kept at precisely controlled temperature of 1270ºC by RF induction heating hence they are placed on Graphite Susceptors. This gives a precise control of Si molecules, in vapour phase of partial pressure P, impinging upon Si Substrate:

(Impingement Rate) F =

$$F = \frac{P}{\sqrt{(2\pi m k T)}} = \left[\frac{3.5\times10^{22}\,P(Torr)}{\sqrt{m(gm)\,T(Kelvin)}}\right]$$

$$\frac{molecules}{cm^2 - sec}$$

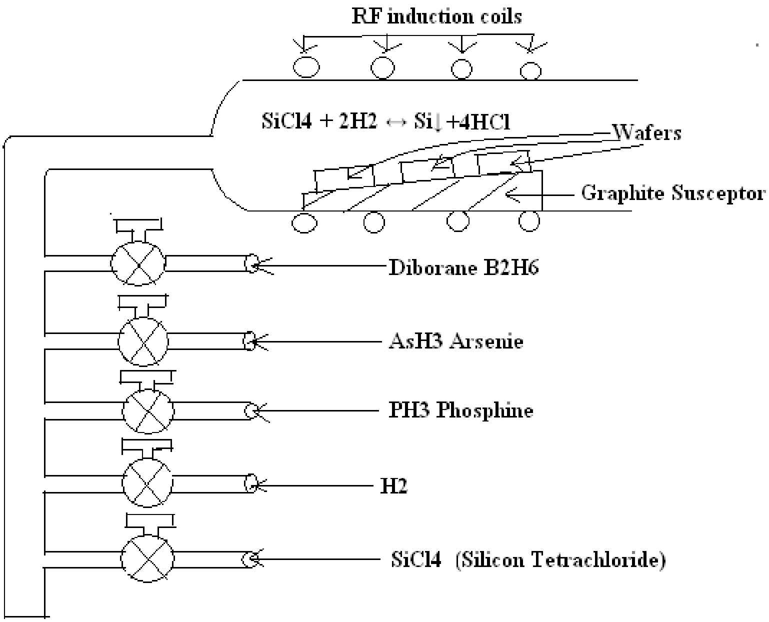
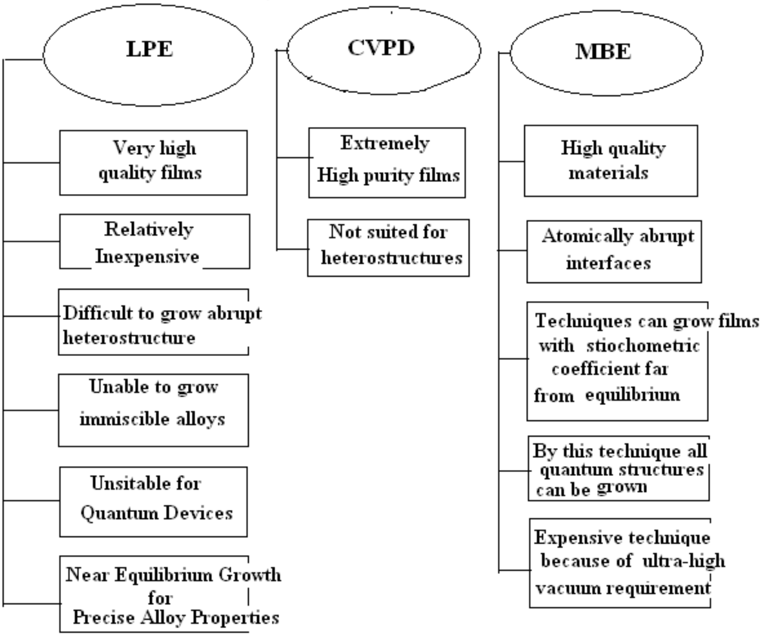


Figure 11. Chemical Vapour Phase Epitaxy for growing epitaxial films of Silicon.

Table 3. Comparative study of LPE, CVPD and MBE



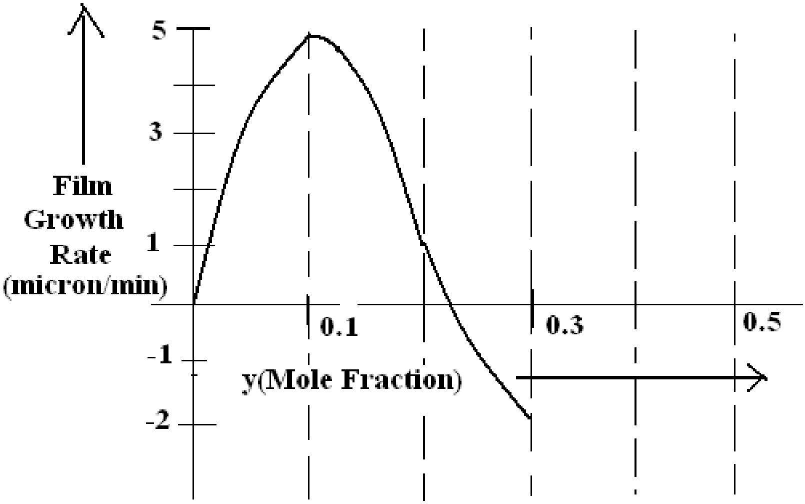
LPE	CVPD	MBE
Very high quality films	Extremely High purity films	High quality materials
Relatively Inexpensive	Not suited for heterostructures	Atomically abrupt interfaces
Difficult to grow abrupt heterostructure		Techniques can grow films with stiochometric coefficient far from equilibrium
Unable to grow immiscible alloys		By this technique all quantum structures can be grown
Unsitable for Quantum Devices		Expensive technique because of ultra-high vacuum requirement
Near Equilibrium Growth for Precise Alloy Properties		

CPVD is suitable for Homo-epitaxy but not for hetero-epitaxy. The byproduct of this reaction is HCl as seen from the chemical equation of reaction:

At 1270ºC $SiCl_4$(Silicon Tetrachloride) + $2H_2 \leftrightarrow Si\downarrow + 4HCl$. 1.

This byproduct HCl can attack the Silicon Substrate and cause chemical etching. Hence Chemical Reaction as depicted Eq.(1) is reversible.

\



Graph 1. Film growth rate vs y(Mole Fraction)

If y(mole fraction = $SiCl_4$ conc./Total conc) is greater than 0.23 as seen from Graph 1 , growth rate is negative. This means chemical etching of Silicon Wafer. The negative growth rate is used for in-situ cleaning of Silicon Wafers. This is one of the ways of chemical etching while preparing Silicon Wafers from Silicon Ingot.

If $y < 0.23$ we have positive growth rate and Si vapour deposition takes place as an epitaxial film. The growth rate is maximum at $y = 0.1$. The maximum growth rate is 5μm/minute.

Since CVPD is carried out at 1270ºC, there is the problem of out-diffusion of Arsenic from the substrate into the epitaxial layer. Arsenic is used as buried layer to reduce the series collector resistance of Vertical NPN transistor. Arsenic has very low diffusion coefficient still at the elevated temperature of 1270ºC, some out-diffusion is inevitable. To prevent this out-diffusion altogether, we use Liquid Phase Epitaxy which is shown in Figure 12.

LIQUID PHASE EPITAXY.

The Liquid Phase Epitaxy set up is shown in Figure 12.This is carried out at 900ºC hence the problem of out-diffusion is completely prevented but Silane which is used for LPE is highly toxic. As seen in Figure 12, there is a super saturated solution of Silane kept at 900ºC. By Graphite Slider, the silicon substrate is brought directly underneath the well containing Silane. As Substrate comes in contact with Silane , latter is dissociated into Silicon and Hydrogen and Silicon precipitates onto the substrate forming an epitaxial film on the substrate.

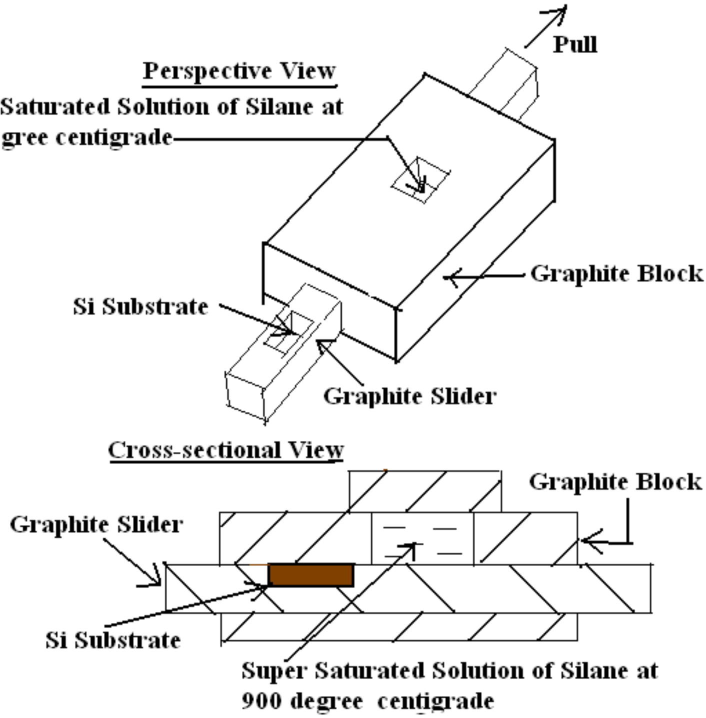


Figure 12. Liquid Phase Epitaxy Setup.

MOLECULAR BEAM EPITAXY

Due to explosive demand of Wireless Communication Equipment there has been a sudden spurt in demand of MBE equipments particularly of multiple wafer MBE equipment which can give a high throughput of epitaxilly prepared wafers.

There are two kinds of epitaxial film growth: **Homoepitaxy** (same composition) and **Heteroepitaxy** (different composition). By Molecular Beam Epitaxy, multi-layered thin films of single crystals of different compositions and of atomic dimension can be grown. In effect we can achieve heteroepitaxy which is the hallmark of compound semiconductor devices such as Photonic Devices and GaAs MESFET.

MBE is a process for making compound semiconductor materials with great precision and purity. These materials are layered one on top of the other to form semiconductor devices such as transistors and lasers. These devices are used in such applications as fiber-optics, cellular phones, satellites, radar systems, and display devices. MBE is used for fabrication of Super-lattices and hetero-junction MESFET. Super lattices are periodic structures of alternating Ultra-thin layers of compound semiconductor.

MBE growth produces complex structures of varying layers which are further processed to produce a range of electronic and optoelectronic devices, including high speed transistors, light-emitting diodes, and solid state lasers. MBE is a powerful technique both for research into new materials and layer structures, and for producing high-performance devices.

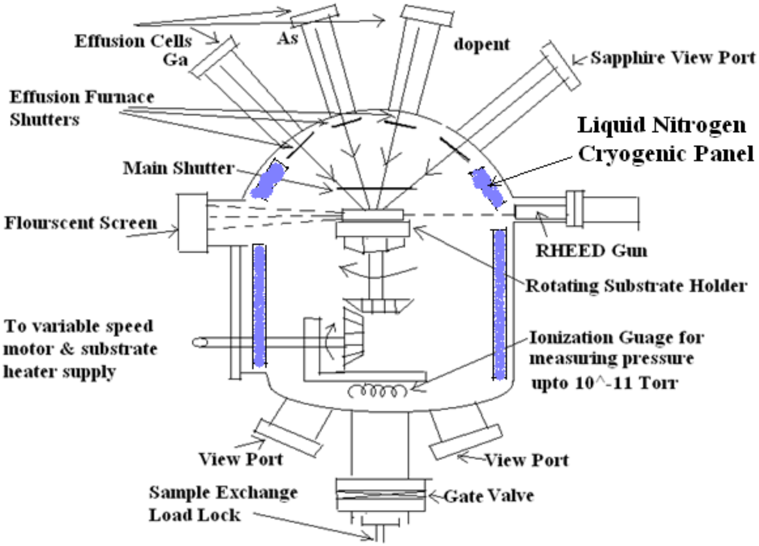

Effusion Cells
Ga
As
dopent
Sapphire View Port
Effusion Furnace Shutters
Liquid Nitrogen Cryogenic Panel
Main Shutter
Flourscent Screen
RHEED Gun
Rotating Substrate Holder
To variable speed motor & substrate heater supply
Ionization Guage for measuring pressure upto 10^-11 Torr
View Port
View Port
Sample Exchange Load Lock
Gate Valve

Figure 13. Molecular Beam Epitaxy Set Up.

Referring to Figure 13:

The walls of the chamber cooled with liquid nitrogen. This cryogenic screening around the substrate minimizes the spurious flux of contaminating atoms and

Sample Exchange Load Lock-this permits maintenance of Ultra High Vacuum while changing the substrate.

Effusion Cells- this contains the solid source to be evaporated and deposited on the substrate. Temperature of the effusion oven is adjusted to give desired evaporation rate.

Rotating Substrate Holder- rotation of substrate ensures less than ±1% doping variation and ±0.5% in thickness variation.

RHEED Gun-Reflection High Energy Electron Diffraction Gun gives a beam of electron which can be made incident on the epitaxially grown film. The diffraction pattern of electrons is studied on the fluorescent screen. This shows a maximum when there is a completed monolayer and a minimum when there is a partial layer as this produces more scattering. Thus RHEED Gun is used for in-situ monitoring of the growth of the epitaxial film mono-layer by mono-layer.

Computer controlled shutters of each furnace allows precise control of the thickness of each layer, down to a single layer of atoms.

Substrate Heating- to obtain high quality epitaxial layer, growth temperature must be relatively high. The substrate wafer must be heated to allow the atoms to move about the surface and reach the proper ordered site. Growth must be at a temperature where growth rate is insensitive of minor temperature variation. Atoms arriving at the substrate surface may undergo absorption to the surface, surface migration, incorporation into the crystal lattice, and thermal desorption. Which of the competing pathways dominates the growth will depend strongly on the temperature of the substrate. At a low temperature, atoms will stick where they land without arranging properly - leading to poor crystal quality. At a high temperature, atoms will desorb (reevaporate) from the surface too readily - leading to low growth rates and poor crystal quality. In the appropriate intermediate temperature range, the atoms will have sufficient energy to move to the proper position on the surface and be incorporated into the growing crystal.

The Inventors of MBE are : J.R. Arthur and Alfred Y. Chuo (Bell Labs, 1960).

MBE is a technique for epitaxial growth of single crystal atomic layer films on single crystal substrate. It gives precise control in chemical composition and doping profiles. To avoid contamination of the epitaxial films, Ultra High Vacuum within MBE chamber is imperative. It requires Very/Ultra high vacuum (10^{-8} Pa or 10^{-11}Torr). This implies that Epitaxial film is grown at slow deposition rate (1 micron/hour). This permits epitaxial growth of single atomic layer if desired. Slow deposition rates require proportionally better vacuum. Ultra-pure elements are heated in separate quasi-knudson effusion cells (e.g., Ga and As) until they begin to slowly sublimate. Gaseous elements then condense on the wafer, where they may react with each other (e.g., GaAs). The term “beam” in MBE means the evaporated atoms do not interact with each other or with other vacuum chamber gases until they reach the wafer. Each gas beam may be turned on and off rapidly with a shutter or a valve. Beam intensity (called the flux) is adjusted for precise control of layer composition. A collection of gas molecules moving in the same direction constitute the molecular beam. Simplest way to generate a molecular beam is Effusion cell or Knudsen cell . Oven contains the material to make the beam. Oven is connected to a vacuum system through a hole. The substrate is located with a line-of-sight to the oven aperture. From kinetic theory, the flow through the aperture is simply the molecular impingement rate on the area of the orifice. Impingement rate is: The total flux through the hole. The spatial distribution of molecules from the orifice of a knudsen cell is normally a cosine distribution. The intensity drops off as the square of the distance from the orifice. Intensity is maximum in the direction normal to the orifice and decreases with increasing θ, which causes problems. Use collimator, a barrier with a small hole; it intercepts all of the flow except for that traveling towards the sample.

During the MBE process, growth can be monitored in situ by a number of methods:

Reflection high energy electron diffraction (RHEED), using forward scattering at grazing angle, which shows a maximum when there is a completed monolayer and a minimum when there is a partial layer as this produces more scattering;

Low energy electron diffraction (LEED), takes place in backscattering geometry and can be used to study surface morphology, but not during growth;

Auger electron spectroscopy (AES), records the type of atoms present;

Modulated beam mass spectrometry (MBMS), allows the chemical species and reaction kienetics to be studied.

Computer controlled shutters of each furnace allows precise control of the thickness of each layer, down to a single layer of atoms.

Intricate structures of layers of different materials can be fabricated this way e.g., semiconductor lasers, LEDs.

Before starting the epitaxial growth, in-situ cleaning of substrate is required. This is achieved by High Temperature Baking of the substrate. This decomposes and vaporizes the oxide layer over the substrate. The second method of in-situ cleaning is low energy ion beam of inert gas is used to sputter clean the surface. After the sputtering , low temperature anneal;ing is required to reorder the surface lattice system.

If there is a lattice mismatch between the substrate and the growing film, elastic energy is accumulated in the growing film. At some critical film thickness, the film may break/crack to lower the free energy of the film. The critical film thickness depends on the Young's moduli, mismatch size, and surface tensions. Hence under heteroepitaxy, we must keep the thickness lower than critical film thickness.

Figure 14 shows the physics of epitaxial growth in MBE system. . The aim of this process is to enable sharp interfaces to be formed between one type of alloy and the next e.g. GaAs and AlAs, and thus create structures which may confine electrons and exibit 2-dimensional behaviour. Molecular Beam Epitaxy (MBE) is basically a sophisticated form of vacuum evaporation.

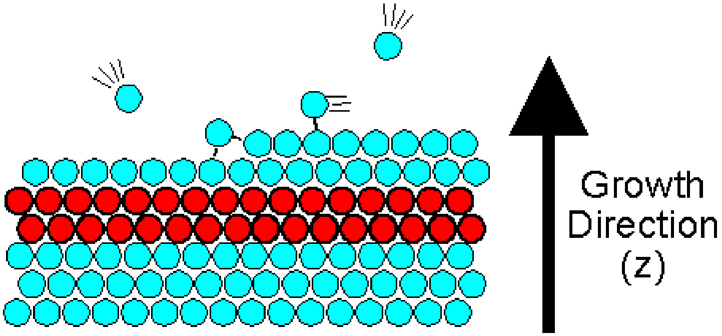


Figure 14. Molecular Beam Epitaxy growth mono-layer by layer or atom layer by layer.

METAL ORGANIC CHEMICAL VAPOUR PHASE DEPOSITION >

The growth process in MOCVD (metal-organic CVD, also known as MOVPE metal-organic vapour phase epitaxy) is similar to MBE, but the atoms are carried in gaseous form to the substrate. GaAlAs growth is achieved by using a mixture of hydrogen as a carrier gas and organometallic precursors such as trimethyl galium and/or trimethyl aluminium together with arsine. The growth rate can be 10 times greater than in MBE, the process does not require ultra high vacuum and it can be scaled up from research to production of commercial devices relatively easily. However, the preparation of the gaseous mixtures has to be very carefully controlled so that as yet it is unclear which technique will eventully dominate. One advantage MOVPE has over MBE is in the ability to grow phosphorous containing alloys, once phosphorous has been introduced into an MBE chamber it is almost impossible to grom anything else! One disadvantage is that in situ monitoring is more difficult.

Syllabus_EC1410A_Solid State Physics and Devices in nano-technology era.

This is the syllabus of 42 hours lecture Course EC1410A[3-0-3]. There are 3 hours lecture and 3 hours Laboratory class per week for 14 weeks in one semester.

Syllabus_EC1410A_Solid State Physics and Devices in nano technology era. [3-0-3] Total Lectures 42.

1. **History of development of Electron Devices from Vacuum Tube Technology to Solid State Technology to NanoTechnology Era-the Harbinger of The Third Wave of Civilization. [2 Lectures]**
2. Semiconductor Physics,Carrier Modelling and Carrier Action:Fermi-Dirac Distribution of Fermions, Intrinsic Semiconductor,Extrinsic Semicnductor, Intrinsic Carrier Concentration, Thermal Equilibrium values of Majority and Minority Carrier Concentration, Law of Mass Action, Compensation, Space Charge Neutrality, Recombination:Shockley-Hall Read Recombination and Auger Recombination, Life-Time of excess minority carriers, mobility, electric drift current, diffusion current, continuity equation in steady state condition. **[6 Lectures]**

3. Classical diodes- Derivation of Ideal Diode Equation/Shockley Equation, Real Diode Equation ,Ideality Factor, temperature dependence of reverse leakage current, temperature dependence of forward bias voltage, Avalanche Breakdown, Zener Breakdown, Backward diode, Tunnel Diode, Varactor Diode, Schottky Diode,difference between rectifying contact and ohmic contact, Junction Capacitance, Diffusion Capaitance, Incremental Circuit Model of a Diode(low frequency and high frequency), Switching Characteristics of a Diode. **[6 Lectures]**
4. Fabrication of Integrated Circuit Vertical NPN transistor,Physics of Operation of BJT: emitter injection efficiency, base transport factor, multiplication factor, base transit time, short circuit forward current transfer ratio (α_F and α_f), short-circuit forward current gain (β_F and β_f) ,transit frequency (ω_T), Incremental Circuit Model of CB BJT (T-Model at low and high frequenies), Incremental Circuit Model of CE BJT(Universal Hybrid-π Model at low and high frequencies). Output Characteristics of CB BJT, Outpt Characteristics of CE BJT(Forward Active Mode, Inverse Active Mode, Saturation Mode, Cut-off Mode), Switching Transient of RTL inverter circuit. **[6 Lectures]**
5. Fabrication of Integrated Circuit n channel JFET(Junction Field Effect Transistor), Physics of Operation nJFET:Ohmic or Triode Region of

operation, Saturation or Pentode Region of Operation, Transfer Characteistics, Output Characteristics(Depletion Type Device or Normally-On device), Symbol of nJFET and pJFET, incremental model of CS nJFET (at low and high frequencies) **[6 Lectures]**

6. Fabrication of Integrated Circuit n channel MOSFET(Metal-Oxide-Field Effect Transistor), Physics of Operation nMOS:Ohmic or Triode Region of operation, Saturation or Pentode Region of Operation, Transfer Characteistics, Output Characteristics, Symbol of NMOSFET , PMOSFET and CMOS, incremental model of CS NMOSFET (at low and high frequencies), Transfer and Output Charateristics of NMOS(E)_enhancement type and NMOS(D)_depletion type.

[6 Lectures]

1. Photonics- direct and indirect bandgap semiconductor, Element and Compound Semiconductor, Band-Gap Engineering, Homo and HeteroJunction Diode, Homo and HetroJunction BJT, Fabrication and Physics of Operation of Photo-Diode, Fabrication and Physics of Operation of LightEmittingDiode(LED), BandGap Engineering of LED to obtain different coloured LED, Fabrication and Physics of Operation of LASER-Diode. **[8 Lectures]**

2. Basic Structure of CHARGE COUPLED DEVICES(CCD) and its application as a CCD Camera in Space Applications. **[2 Lectures]**

History of Electron Devices- Vacuum Tube Era, Solid State Era and NanoTechnologyEra(Untitled)
Section 2.1.1 describes the historical development through Vacuum Tube Era.

2.1. History of Electron Devices- Vacuum Tube Era, Solid State Era and NanoTechnologyEra.

In 1897, while studying Cathode Ray in a Vacuum Tube, J.J.Thomson for the first time identified the Cathode Ray with ‘Electrons’. For this J.J.Thomson was awarded Nobel Prize in Physics in 1906.

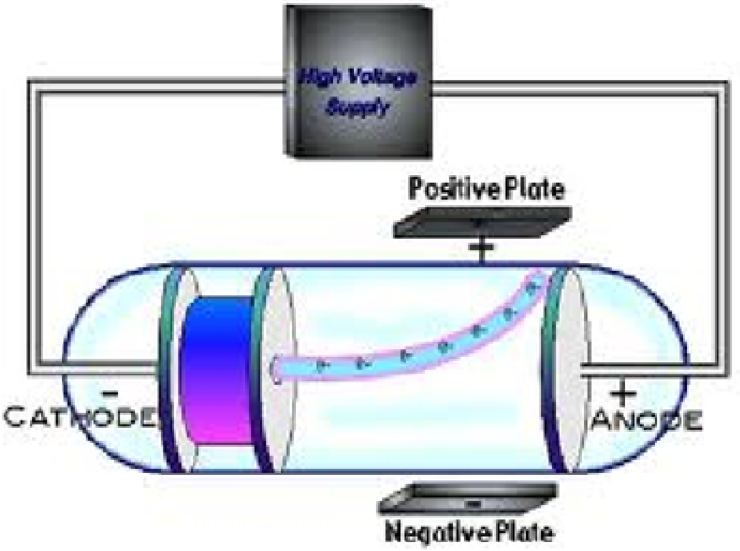


Figure 2.1. Deflection of Cathode Ray towards positive plate while travelling towards Anode in a vacuum tube.

Thomson’s Vacuum Tube was a very simple contraption. The Tube was evacuated to a vacuum of 10^{-3}Torr by a rotary pump. 1 Torr is 1mm Hg Vacuum. 1 Atmosphere pressure will support 76cm column of Hg. We can go to still higher vacuum using diffusion pump. For MOSFET fabrication we require ion-pump which can evacuate to 10^{-8}Torr.

This evacuated Vacuum Tube (VT) contains a filament, cathode and anode. Filament is made of tungsten which on passage of current gives a lot of thermal energy by resistive heating. This heats the Cathode.

Cathode is made of Strontium-BariumCuO with a work-function WF = 1eV.

The energy band diagram of Cathode is given in Figure 2.2.

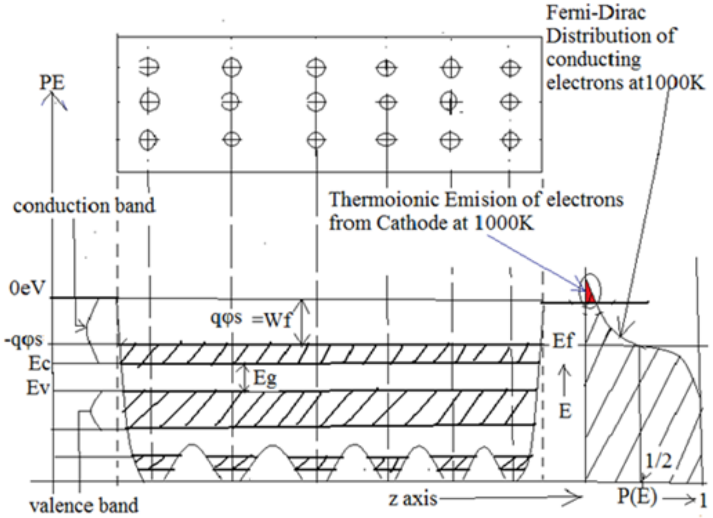
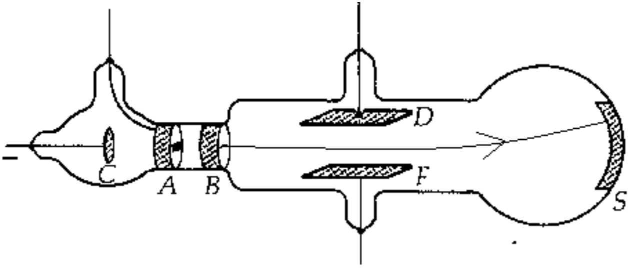


Figure 2.2. Energy-Band Diagram of Metallic Cathode and thermo-emission of electron at 1000K. The tip of the skewed Fermi-Dirac distribution above 0eV , red colour, gives the probability of Thermo-ionic emision of electron out of the cathode in the vacuum surrounding the cathode.

As seen in Figure 2.2., Cathode is heated to 1000K.Conducting electrons in the Cathode are energetic enough to cross the Surface Barrier Potential (φs) and constitute the thermoionically emitted electrons which form a space charge cloud around the cathode as shown in Figure 2.4. As emitted electrons leave the surface of the cathode, a positive charge builds up on the cathode surface because of unbalanced charge within the cathode. This positive charge holds on to emitted electrons causing space charge cloud. From this space charge cloud electrons are pulled to the anode held at positive potential of 100V. This motion of electrons from the cathode to anode constitute the Anode Current and which was interpretated as Cathode Ray until Thomson revealed the mystery of thermoionic emission of electrons from the cathode.

Thomson studied this beam of electron meticulously and established the charge ‘q’ and q/m ratio of electron where m is the mass of electron. The set up is shown in Figure 2.3.



Thomson's tube for measuring q/m for the particles of cathode rays (electrons). Electrons from the cathode C pass through the slits at *A* and *B* and strike a phosphorescent screen *S*. The beam can be deflected by an electric field between plates *D* and *F* or by a magnetic field (not shown).

Figure 2.3.Thomson set-up for measuring q and q/m of electron

The verification of the negative charge of electron led J.J.Thomson to hypothesize the Plum-Pudding

Model. He suggested that negatively charged electrons are swimming in a sea of positive charge. It was only later on that by Gold-foil experiment Rutherford decisively established the planetary model of atom. He found that 99% of the mass of the atom is concentrated in the nucleus and it was positively charged. Electrons are orbiting this nucleus and their opposite charges are equal in magnitude to make an atom neutral in charge.

The discovery of electrons formally launched the Branch of Engineering we know as Electronics and Communication Engineering. This was also the start of Vacuum Tube Era.

2.1.1. Vacuum Tube Era

Vacuum Diode - After the discovery of electron in 1897, Vacuum Diode was invented by John Ambrose Fleming in 1903. Vacuum Diode is identical to the CRT shown in Figure 2.1 except that it does not have parallel set of deflecting plates and is shown in Figure2.4. I-V characteristics of the Vacuum Diode is shown in Figure 2.5.

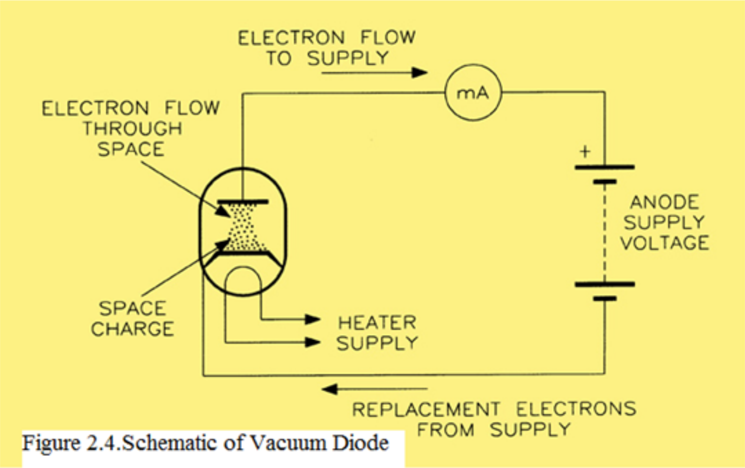


Figure 2.4.Schematic of Vacuum Diode

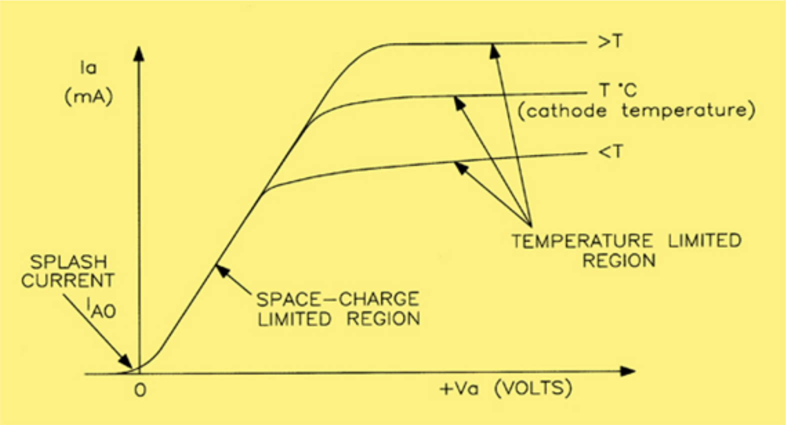


Figure 2.5. I-V characteristics of Vacuum Diode under forward bias.

The Physics of current flow in Vacuum Diode is as follows:

1. By Thermo-ionic emission, a space charge cloud is formed surrounding the cathode.
2. As Anode is made positive, electrons are drawn towards Anode. As long as number of electrons

drawn by Anode is less than electrons emitted from the Cathode, we are in Space-Charge Limited Region and Anode Current (I_A) increases linearly with Anode Voltage(V_{AK}) according to 3/2 Power Law as given in Equation 2.1.

3. As soon as electrons drawn by Anode are equal to electrons emitted by Cathode, we enter Temperature-Limited Region. In Temperature Limited Region, Anode Current saturates for a constant temperature. If temperature of Cathode is increased, the saturation value of Anode Current will increase given by Equation 2.2.The Work Function of Cathode is vital in the Temperature Limited Region. At 2500ºC, Tungsten ($W_F = 4.52$eV) cathode gives thermo-ionic current density of 3000A/m^2 and SrBaCuO ($W_F = 1$eV) cathode gives 7.34×10^9A/m^2. There is almost 6 orders of magnitude improvement in thermo-ionic emission. Hence with SrBaCuO we can operate at 800ºC comfortably with significant thermoionic emission.
4. At zero Anode Voltage current in the circuit should be zero but it is not. Because of statistical fluctuation in the velocity, some electrons are energetic enough to reach the Anode at 0 V even. This is known as splash current.
5. Under Reverse Bias, there is no reverse diode current. But at very high voltages by quantum

mechanical tunneling electrons can be pulled out from the anode to cathode and a phenomena similar to break down occurs in Vacuum Diode also. This is called high-field emission.

$$J(space\_charge\_limited) = \frac{K}{d^2} \times V_{AK}^{3/2} \left(\frac{mA}{mm^2}\right) \_\_\_\_\_\_ (2.1)$$

$$Here\_K = constant = \frac{4}{9}\varepsilon_0 \left(\frac{2|q|}{m}\right)^{1/2} = 0.002334\left(\frac{mA}{V^{3/2}}\right)$$

$$V_{AK} = potential\_difference\_betwen\_Anode(A)\_and\_Cathode(K)$$

$$d = distance\_between\_A\_and\_K.$$

$$J(Temperature\_Limited\_Region) = A_0 T^2 Exp\left[-\frac{W_F}{kT}\right] \_\_\_\_\_\_ (2.2)$$

$$Where\_W_F = work\_function\_of\_cathode = 1eV(SrBaCuO) \& 4.52eV(Tungsten)$$

$$A_0 = \frac{4\pi qmk^2}{h^3} = 1.2 \times 10^6 \left(\frac{A}{m^2 K^2}\right);$$

$$T = temperature\_of\_cathode\_in\_Kelvin;$$

Vacuum Triode - In 1905, Lee de’Forest invented Vacuum Triode. It had three electrodes: Cathode, Grid and Anode. Grid was the third electrode inserted between Cathode and Anode. This controlled the Anode Current. For zero Grid Voltage, anode current flew at zero Anode Voltage onward just as in Diode.. As Grid Voltage became more negative, anode current onset was delayed and diode I-V characteristic was displaced as shown in Figure 2.6(b). In effect Triode behaved as Voltage-Controlled-Voltage –Source(VCVS). Lee de’Forest used this three terminal device to assemble RC-coupled Audio Amplifier the enabling technology for long-distance telephone call. Telephone had been invented in 1876 by Alexandre Grahm Bell but it could not become Public Utility Service until 1913.In Figure 2.6(a) a RC-coupled Audio Amplifier is shown.

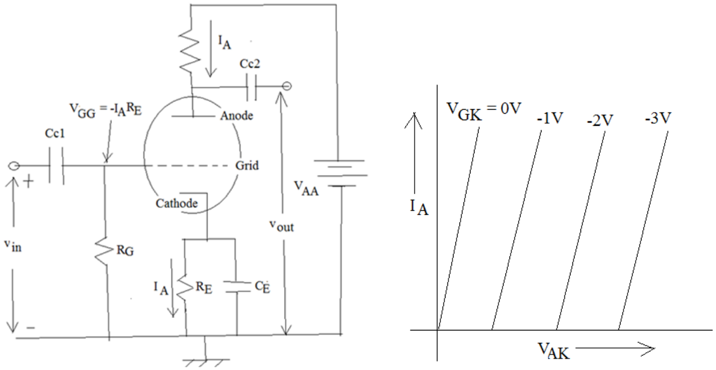


Figure 2.6. (a) Vacuum Triode connected as Common Cathode RC-coupled Amplifier;(b)Triode I-V characteristics.

In quick succession Tetrode and Pentode were invented.

Tetrode and *Pentode*- Tetrode had four electrodes . The fourth electrode was a screen electrode held at about +70V. The Tetrode and its output characteristics are shown in Figure 2.7.

As seen in Figure 2.7, we get parallel set of horizontal lines akin to Voltage Controlled Current Source but with a wiggle. This wiggle is due to “secondary emission of electrons” from the Anode. As the highly accelerated electrons, due to the positive screen voltage, strike the anode it causes local thermal heating which leads to secondary thermo-ionic emission. This causes the wiggle which leads to non-linear distortion and reduction in gain.

To suppress this secondary emission, a fifth electrode at negative voltage with respect to anode is introduced. This repels back the secondary emission and removes the wiggle. This fifth electrode is known as suppressor grid connected to the cathode. This tube with five electrodes is known as Pentode and shown in Figure 2.8. This device approximates an ideal Voltage-Controlled Current Source.

These topics have been discussed in the introductory chapters of Lectures on Analog Electronics (Collection 11167 uploaded on cnx.org). The I-V characteristics of Triode and Pentode are shown in the module, m29634, and these are equivalent to controlled sources namely Voltage-Controlled-Voltage-Source and Voltage Controlled-Current source respectively. These controlled sources are discussed in the module, m33375, in the same collection of Lectures on Analog Electronics.

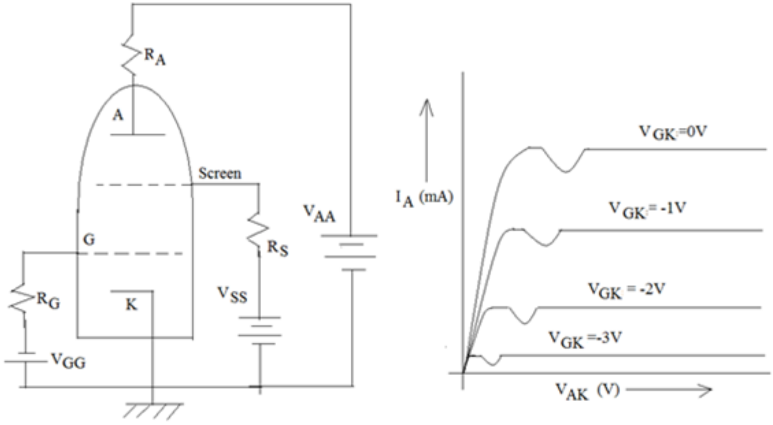


Figure 2.7. Schematic of a Tetrode and its output charateristics.

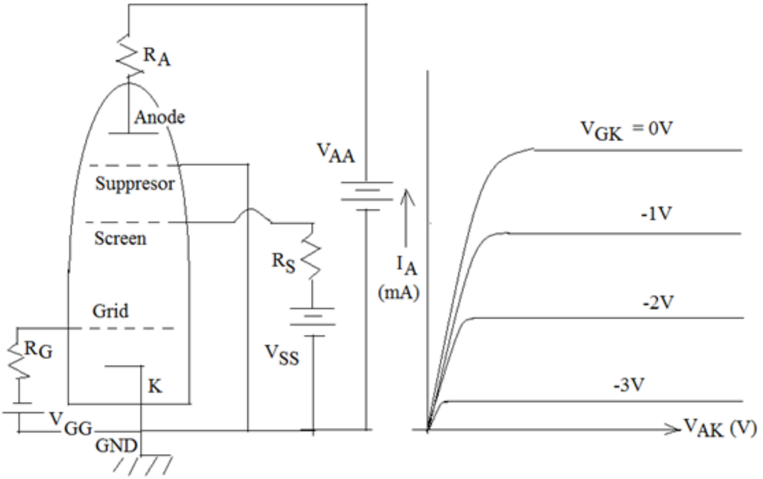
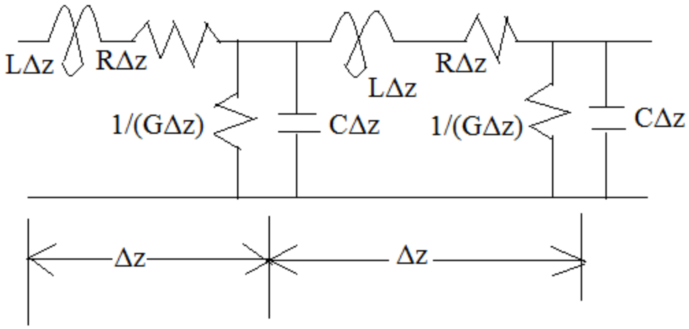


Figure 2.8. Schematic of Pentode and its output characteristics.

2.1.1.1. RC-coupled Amplifier and Heaviside Distortion-less Transmission enables the long-distance Telephone Call across N.American Continent.

In 1911, the first long distance call was successfully made from New York to Denver, Colorado. No Amplifiers had been used in this 2000miles long distance telephone call. Heaviside Distortion-less Transmission condition had been satisfied to enable this call.

Oliver Heaviside was deeply involved with telegraphic transmission and its theorization. He treated a long transmission line as distributed parameter network as shown in Figure 2.9.



R=series resistance per Km of line;L=series inductance per Km; C=shunt interwirng capacitance per Km;G=shunt conductance per Km; Δz=length of the elemental section of the transmission line.

Figure 2.9. Distributed Parameter Transmission Network.

By circuit analysis he showed that distortion-less transmission requires that:

$$\frac{G}{C}=\frac{R}{L} \qquad (2.3)$$

But in real life transmission telephonic cables, this balance is not achieved. The actual imbalance is:

$$\frac{G}{C} \ll \frac{R}{L} \qquad (2.4)$$

This imbalance results in serious propagation-delay distortion or dispersion of the signal. This result is phase-distortion and the voice gets jumbled up in long distance telephonic call. To correct this, Oliver Heaviside suggested Inductive-Coil loading of the transmission line and thereby achieve distortion-less transmission condition. In this remedy, single high-induction coils are connected to the transmission line in series at regular intervals. This enabled the clear reception of a call coming across 2000 miles without any kind of amplification.

In 1913, periodic inductive loading as well as RC-Coupled Amplifier Repeater stations were used at intervals of 100Km distance. This enabled a clearly audible long-distance telephonic call from New York to San-Francisco, California, over a distance of 4000 miles. This trans-continental telephonic call set the stage for turning Telephone into Public Utility Service. This laid the ground work for the Information Revolution.

Because of the poor quality of telegraphic cables laid down under Atlantic and Pacific Ocean, trans-Atlantic or trans-Pacific Telephone calls still could not be made. These trans-Oceanic Calls were put through Radio-Telephony. Only in 1956 when first Voice-quality trans-Atlantic Coaxial Cables,TAT-1, were laid down that Trans-Atlantic Telephone Calls became a reality.

In 1945, ENIAC (Electrical Numerical Integrator and Calculator) using 18,000 Thermo-ionic Vacuum Tubes was built up. It consumed 200kW power, it weighed 27 Metric Tonnes and it occupied a large size Hall. At the time computers were hardly user friendly.

2.1.2. Solid State Era

Section 2.1.2. Gives the historical development of Solid State Devices up to nanotechnology era.

2.1.2.Solid State Era of Electronic Devices.

In 1949, in Bell Laboratories Shockley, Bardeen and Braittain invented the first Solid State Three Terminal Device called Bipolar Junction Transistor (BJT). Two terminal solid state diodes were already in use. Using this BJT, potable, battery operated transistor Radios (AM Receiver Sets) were launched. This brought a mini-revolution in Entertainent Electronics. Light weight, cheap, micro-miniaturization and portablity was going to be the hall mark of the Solid State Era. The communication infrastructure will be made compact, less power hungry, reliable and integrated in the services (audio + video + data) it could provide. The computers will be transformed into personal computers. Mobile smart phones will become Palm-Top Computers. This is what we call *Third Infortmation Revolution1*.

1[**First Information Revolution** was the invention of Pheonician Alphabets in the West in 700BCE , the invention of Brahmi Lipi in East in 400BCE and the invention of Kharosthi Alphabets which was derived from Aramatic Script used in Persia earlier than 300CE. Pheonician Alphabets became the basis of all European Languages. Brahmi Script became the basis of all the Indian languages namely Telegu, Tamil, Kanad, Malyalam, Bengali, Uria, Assamese, Gujrati, Maharashtri, Gurumukhi and Devnagri. Kharoshti Scripts and Aramatic Scripts became the basis of Arab, Persian and Urdu written from right to left.

Second Information Revolution was the invention of Rotating type Printing Presses in 1436. This is also known as Gutenberg Revolution. Johannes Gutenberg, a goldsmith and businessman from the mining town of Mainz in southern Germany, borrowed money to invent a technology that changed the world of printing. Johannes Gutenberg invented the printing press with replaceable/ moveable wooden or metal letters in 1436 (completed by 1440). This method of printing can be credited not only for a revolution in the production of books, but also for fostering rapid development in the sciences, arts and religion through the transmission of texts. Till then hand written Manuscripts on parchements, Palm-leaf or papurus was in use which allowed only a selective group of people to access the written , scholarly text. But printing books and pamphelets changed this limited accessibility to written information.

It was printing and rapid diffusion of printed material which brought the Reformation Movement by Martin Luther King in Germany. This led to the establishment of Protestent Church in 1517-21 CE. This led to the overthrow of Monarchy by Jacobin Party in France in 1789 CE. This led to the American Independence under the leadership of George Washington of Democratic Party on 4th July 1776 CE. This led to the unification of Germany under the Junkers in 1871 CE and unification of Japan by the Meiji Oligarchs in 1867 CE.

In nut-shell, Second Information Revolution laid the foundation for Industrial Revolution where we will have Capital-intensive Enterprises.]

2.1.2.1.Main Frame Era.

In 1956, Sperry Rand,USA, commercialized 2nd Generation Main Frame Computers by the brand name of UNIVAC II. This was a hybrid of Vacuum Tube (5,200 Tubes) and BJT (1,200 Transistors) hence size and power consumption was reduced but it did not carry the distinct hall mark of Solid State Era. But it was a great improvement over ENIAC in terms of weight, physical size and power consumption. Also it was Storage Program Computer whereas through Plugboard Control ENIAC had to be manually adjusted for running different programs.

On 13th September 1956, IBM 305 RAMAC

Computer with disk-drive was commercialzed. It was still a Vacuum –Tube version but it had secondary storage space of 4.4MegaBytes on 350 disk file. This 350 disk file had 24″ magnetically coated platters 50 in number. Today we can store 70GB information in Hard Disc of computers.

1959 was the turning point in Solid State Era. Robert Noyce of Fair-Child and Jack Kilby of Texas Instruments simultaneously invented the magic chip called Integrated Circuits Micro-Chip. This became the enabling technology of our present on-going Computer-Communication Revolution. This became the engine of growth of Semiconductor Industry. This IC Chip made electronics, electronics communication and electronics computation accessible to one and all. IC Chips made application of Electronics Systems adaptable and suitable for all domains of life. IC Chips are being used by Military, Trade & Commerce, Entertainment, Transportation, Health-care , Industrial Manufacturing , Agricultural production and mining. The list goes on. Every day brings a new domain under the perview of Electronics IC Chip.

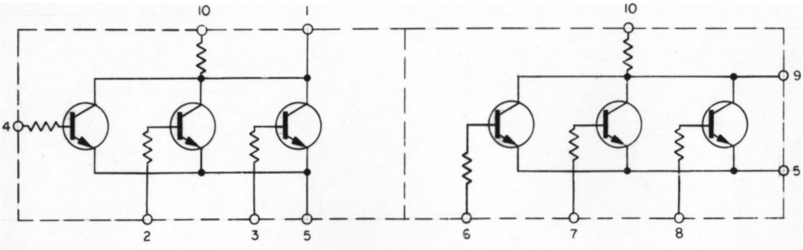


Figure 2.10. Schematic of Dual-3-input-wide-RTL-NOR-Gate used in Apollo Guidance Computer.

In Figure 2.10, the schematic of Dual-3-input-wide-RTL-NOR-Gate is shown. This has been integrated on a Si-die, mounted on a header, gold-wire bonded to the 10 leads of the header by Thermo-compression Bonding and hermetically sealed. The Si-die on which the circuit has been integrated is shown in Figure 2.11. In Figure 2.11, the top-view of the Planar Circuit is shown. It shows the transistors, its interconnections and bonding pad layout. The golden wires bonded to the bonding pads of the circuit and to the external leads of the header are also shown. The external leads of the header are not visible. In Figure 2.12, manually drawn layout of the circuit is shown. It is identical to the layout in Figure 2.11. Thermo-compression bonding of the gold wire to the bonding pads on the Si-Die and to the external leads of the header are shown in Figure 2.13 and hermetic shielding and packaging are shown in Figure 2.14. We have two kinds of packaging: Dual-in-Line Package (DIP) and TO-5 Package. For low power dissipation Ceramic DIP is used whereas for large power dissipations, metal can TO-5 package is more suitable and it can be mounted on a heat sink to prevent thermal run-away condition.

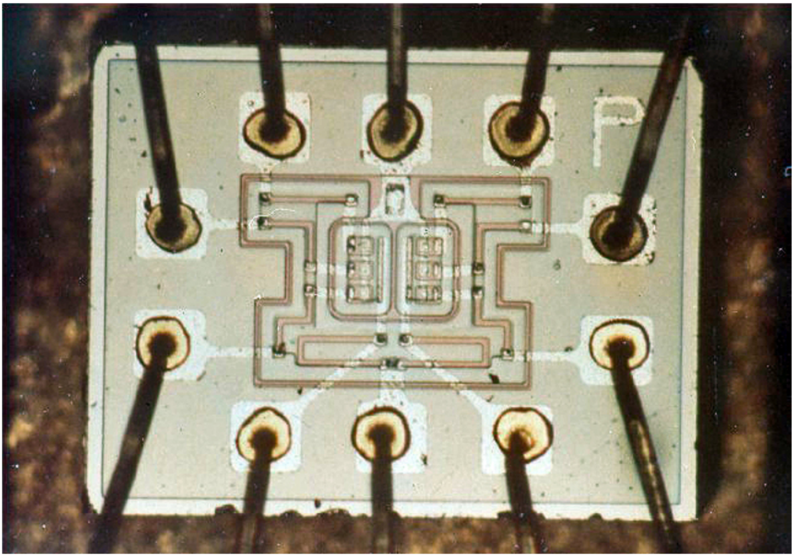


Figure 2.11. Integration of all the components of RTL NOR Gate on one Silicon Die.

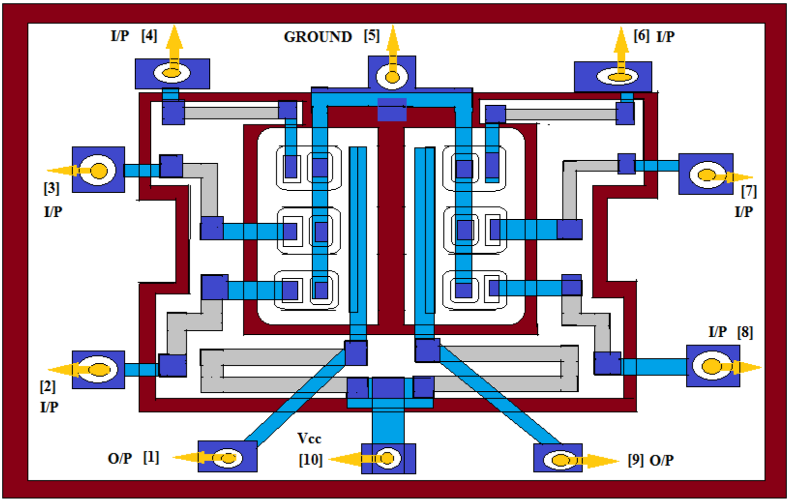


Figure 2.12.Manually drawn, several times magnified, layout of the Monolithic Circuit of Dual-3-input-wide-RTL-NOR-Gate, interconnections and the bonding pad.

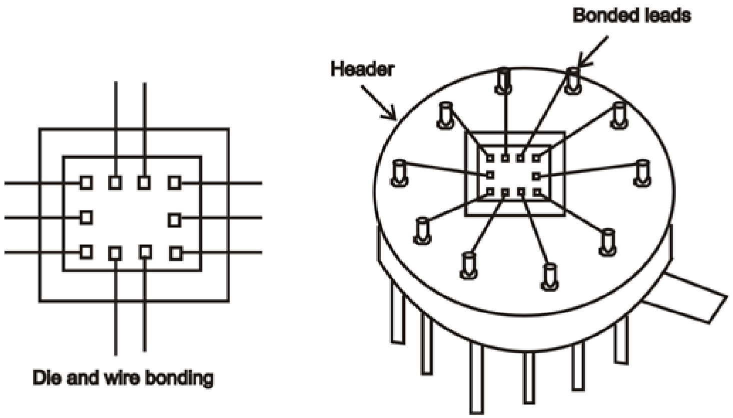


Figure 2.13 **The individual die or chip mounted on the header with leads bonded in place.**

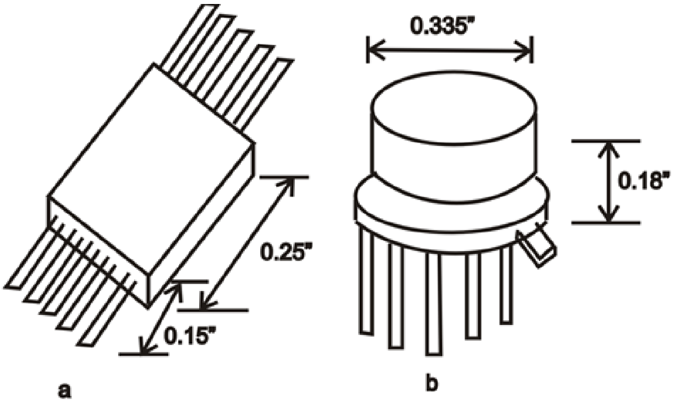


Figure 2.14. Comparison of (a) Dual-in-Line Package(DIP) and (b) 10 lead TO-5 Package.

5th October 1959, the legendary IBM 1401 Data Processing System was introduced. A very good description of this main frame computer is given in the article :

“The Legendary IBM 1401 Data Processing System” by Robert Garner and Fredrick (Rick) Dill, IEEE

Solid State Circuits Magazine , Winter 2010, Vol.2, No.1, pp.28-39.

It was an all transistor Main Frame Computer. Using Automatic Logic Design Software, the volume and power were reduced by 50% and 90% respectively. Standard Modular System of packaging was used which made it compact. The Logic Family used was Saturated-Transistor-Transistor-Logic (TTL-74 Series). The clock rate was 250kHz. Compare this with todays eight core system which has a clock rate of 3GHz in 2013. Memor-to-Memory Architecture was used This reduced the number of instructions needed for acconting and business programs. IBM 1401 had a entry level rental price at US $2500 per month as compared to the previous rental price of $40,000 per month. This economy in rental allowed IBM to capture 50% of Office Computerization.

Full size system had 500,000 discrete components, weighed 4 tonnes, consumed 13,000W, used 10,600Ge-alloy-junction BJT, used 13,200Ge-point contact diodes, it had 16kBytes memory space and it clocked 87kHz.

In 50 machine cycles i.e.. in 0.5ms the computer could add 2 positive 20-digit numbers. Today Pentium IV with a clock rate of 2.5GHz requires only 20ns. 50 machine cycles*(1/Clk_Rate)= addition time of 2 positive 20 digit numbers.

Using the above formula we get the computation time. Clock rate of 87kHz corresponds to 11.5μs Machine Cycle time. Hence 50 machine cycles will take 0.57ms.

If we assume a Clock Rate of 2.5GHz then we have a Machine Cycle Time equal to 0.4ns. So a addition which takes 50 Machine Cycles will take a computation time of 20ns.

Magnetic Core-Memories were not solid state hence they were bulky and 300 million times more costly tham DRAM. The speed and print quality was very good hence IBM 1401 became the Industry’s work-horse. It contributed to the growth of programming as a profession and to the growth of software as an Industry. Use of IBM 1401 in offices created a huge demand of Programmers as well as Software Engineers. But it had a problem with standardization and compatibility. Hence by 1964, IBM 1401 was completely replacxed by S/360 which was compatible and scalable computer family.

2.1.2.2. Mini-Computer Era.

In 1960, Digital Equipment Corporation(DEC) introduced PDP-1(Programmable Data Processor). It was Cabinet Size Office Computer at a total cost of $120,000.00. It was all-transistor System with magnetic-core-memory. Cloak Rate was 200kHz hence Machinew Cycle Time was 5μs. It performed at 100,000 Operations Per Second(OPS).

One operation or instruction took 2 machine cycles hence it took 10μs. Therefore number of Operations Per Second = 1/10μs = 100,000 OPS.

This was the first computer which introduced the concept of user interactive. It created hacker’s culture at MIT. It allowed a game ‘SpaceWar’ to be played.

By 1961, first RTL digital IC chip was commercialized and in 1964 firstr analog IC Chip MC1530 (Operational Amplifier) was commercialized.

In 1964, DEC introduced PDP-8 using IC Chips hence it had very distinct hall-mark of Solid State Era. It was a Table Top computer. Cost came down to $18,000. Clock rate was 667kHz and it had 4kB memory upgradable. Its machine cycle time = 1.5μs and one instruction took two machine cycle times hence $1/(2\times1.5\mu s) = 0.333$MIPS. Therefore this Table-Top computer had 0.333Mega Instruction per Second processing power.

In 1965, Gordon E.Moore submitted a paper on the trend of packing density of components in an IC Chip. [“Cramming more components onto Integrated Circuits”, Gordon E. Moore, *Electronics Magazine*, April 1965, pp.4]. The paper noted that the number of components in integrated circuits had doubled every year from the invention of the integrated circuit in 1958 until 1965 and predicted that the trend would continue "for at least ten years". His prediction has proven to be uncannily accurate, in part because the law is now used in the semiconductor industry to guide long-term planning and to set targets for R & D.

1968 became another water shed in Electronic Device gradual evolution. Robert Noyce of Fairchild, Gordon Moore of Fairchild together with Andy Grove also from Fairchild established INTEL(Integrated Electronics) which was going to become the pace setter of Semiconductor Industry in the coming decades and it continues to play that historic role.

In December 1968, Fall Joint Computer Conference, San Francisco, Douglas C. Engelbart(1925-2013) demonstrated On-Line fully interactive Computer System. He sat on a stage infront of a mouse, a keyboard and other controls and projected the Computer Display on a 22ft high video screen. He demonstrated how a networked interactive computing system would allow information to be shared rapidly among collaborating Scientists. He demonstrated how a mouse which he invented four years earlier could be used to control a computer.

In 60s, mammoth unwieldy Main Frame Computer existed where a deck of cards had to be punched and fed into the computer. After a couple of hours of wait, the print-out of the result would be available. This why a Networked Interactive Real-Time Computing System was REVOLUTIONARY. NLS(the online system) allowed researchers to share information seamlessly and to create and retrieve documents in the form of structured electronic library.Text editing, video conferencing, hyper-text and windowing were demonstrated. This Technology was refined at Xerox's Palo Alto Research Center(PARC) and at the Stanford Artificial Intelligence Laboratory. APPLE and MICROSOFT went on to commercialize and practically brought the Information Revolution in our Modern Life.

National Medal for Technology and Innovation in 2000CE was awarded to Douglas.This Award was for creating the foundation of personal computing including continuous real time interaction based on CRT displays and the mouse, hypertext linking, text editing, on-line journals, shared screen tele-conferencing and remote collaborative work. He created the personal computing component of the Computer Revolution.

2.1.2.3. Microcomputer Era- the beginning of Personal Computer Revolution.

In 1971, Intel marketed micro-processor μP4004 – a 4-bit central processing unit.It combined the function of 12 subsystems on one single Si-die. It was 1cm × 1cm die with 2300(E)PMOS employing 10μm-process Technology which cost $299 and in India it could be obtained for Rs 7,500. It had a clock rate of 184kHz. Assuming 2 machine cycles per instruction it could carry out 92,000 Instructions Per Second. In 2010 Intel is accommodating 560Mega CMOS employing 32nm-process technology and second generation high K-metal Gate Si- technology.

In 1974, IBM's Robert Dennnard, inventor of single transistor DRAM(Dynamic RAM) showed that when MOSFET geometries, voltages and dopings were scaled down, gate transit time also scaled down and performance thus improved by the same factor. ["Design of Ion-Implanted MOSFET with very small Physical Dimensions", *IEEE Journal of Solid State Circuits,* Vol.SC-9,No.5, 256-268,October 1974]. Taken together, density improvement as predicted by Moore's Law and the performance improvement by Dennard signaled a coming explosion of growth in chip processing power. This precisely happened and this is what has been fueling Computer-Communication Revolution till date.

In 1976, Xerox PARC(Palo Alto Research Center) and CALTECH,Pasadena, launched a joint collaborative programme on Electrooniuc Design Automation(EDA) to deal with the complex task of future chip designing where integration level will

exceed 100,000 components. By 1979, Lee Conway and Carver Mead invented the Scalable Design Rules and Structured Design Methodology from top-to-bottom. Their work was incorporated in ‘The introduction to VLSI Systems’ published by Addison-Wesley. Soon after MPC79 programme was launched coordinated by ARPANET.
[“Reminiscences of the VLSI Revolution- How a series of failures triggered a paradigm shift in digital design”, Lee Conway, *IEEE Solid State Circuits Magazine,*Fall 2012, Vol.4, No.3, pp.8-32].

Under MPC79, 12 Universwities participated with LSI Design Courses being offered at Graduate level in Fall Semester. In all 82 design projects from 124 designers spread across 12 die-typesa on two wafer sets were completed. Turnaround time from design cut-off to distribution of packaged chips was only 29 days. At second VLSI Conference in January 1980 the detailed results of MPC79 were presented.

DARPA (Defense Advanced Research Project Agency) seeing the potential of VLSI Design Methods allocated $50 million to VLSI Research. PARC MPC system Technology and Methods of Operation were transferred to University of South Carolina-ISI (a major DARPA software contractor). ISI announced MOSIS service. It is a Multi-Project Wafer Integrated Circuits Service Provider- one of the oldest foundries. Universities and small companies had access to state of art digital technology. On the way it may be mentioned that Advanced Computation System at IBM had pioneered the super-scalar architecturte of µPs and all high performance µP are based on this architecture.

In 1977, a complete stand-alone Personal Computer by the name APPLE II was launched at West Coast Computer Fair. It used MOS 6502µP. It had 1MHz clock rate, 4kB memory and Integer Basic (its driving software) written in ROM.

In 1981, IBM introduced PC5150 based on Intel 8088µP.It had a memory of 40kB RAM. It had floppy disc drive for loading the Operating System.

Multi-Core Processors: By scaling we improved processor speed but simultaneously over-heating problem was caused. Hence instruction level parallism was introduced. This is known as Superscalar Architecture. Quadrupling the physical resources led to linear increase in computational performance. In 2005, Intel marketed first dual-core Pentium-D. Today eight-core Pentium 4th Generation Intel Core i7 Processor uses hyperthreading technology. This technology allows each processor core to work on two tasks at the same time improving multi-tasking, speeding up the work flow and accomplishing more in less time.

Supercomputers and present day multi-core

Personal Computer: The fastest supercomputer has reached the top speed of 10 Peta FLOPS(10×10^{15} Floating-Point Operations per Second). These machines can solve huge complex problems that a PC cannot even touch. Some of the problems solved by supercomputers are: Weather Fore-casting, Complex problems of Quantum Physics, huge simulations pertaining nuclear chain reaction. These super-computers are room-sized and expensive. But the advancements in Solid State Technology has allowed present day PCs also to work at a fraction of Supercomputer's Computational Power. Adapteva has developed Parallela Platform which can run at 90GFLOPS. Intel 750 can run at 9GFLOPS.

Foundaries and Fabless Companies: Very early in IC chips design and fabrication an innovative business model was developed by a small group of people which became the mainstay of IC industries.

Traditional Business Model was Integrated Device Manufacturing. Intel, AMD and Samsung follow this model.The Design and Fabrication was done under one roof. But in the new model, Fabless Company focuses on concept, development, design, sales and marketing.The Fabless Companies are:Qualcomm, Broadcomm,Nivida and Apple. The design is sent to the Foundaries which carry out the actual fabrication in bulk – mask making, diffusion, metallization, splitting the dies, mounting the dies on headers, bonding the die pond pads to the external leads and packaging. The ready packages of IC Chips are sent back to the Parent Companies who will do the actual marketing.

The few Foundaries in the World are:Taiwan Semiconductor Manufacturing Company (TSMC), Global Foundaries in UAE, IBM foundary in USA, Samsung Semiconductor in South Korea, SMIC in China, Tower Jazz in Israel, SSMC in Singapore and X-Fab in Germany. Unfortunately we in India had Semiconctor Complex Limited in Chandygarh but it could not grow into a full scale foundrey.

The top Fabless compnies are QUALCOMM(USA), BroadCom(USA), AMD(USA),NIVIDA(USA) and Marvell(USA). Today the Foundaries and Fabless Companies are responsible for 5 Trillion Dollar worth of World Production which is 1/10 of total World Production of Merchandise of worth $50 Trillion. These are Semiconductor chips, its related systems and services. Semiconductor Industry has become the main engine of growth of the World Economy.

Computers have been the top application for ICs since they were first invented in the late 1950s. But PC market dynamics are changing, in part because of competition from mobile devices like smartphones and tablets.Market research firm IHS iSupplier recently projected that shipments of PCs will contract this year for the first time since

2001.The tremendous growth in smartphone sales over the past five years,coupled with the rise of 4G LTE (Long Term Evolution),have reshaped the cellphone chip market landscape,with Qualcomm Inc. and Samsung Electronics Co. Ltd. capitalizing the most, according to market research firm IHS iSuppli.

2.1.2.4. The Internet Revolution.

In 1982, French Telecom launched MINITEL system. This was the first On-line Public Utility Service. It was the first pre-World-Wide-Web service. Customers could log-in and conduct various search and commercial activities.

On 1st January, 1983, Standardized TCP/IP rolled out formally and it was incorporated in ARPA-sponsored packet networks and in Local Area Networks.

In 1989, hyper-text-transmission protocol (http) is introduced as a Website address. Web browser is introduced the same year.

By 1991, World-Wide-Web was introduced on internet without royalities. Internet is a vast network of computers, interconnected in different ways but they all speak the same language namely Hyper Text Mark Up (HTML)language. WWW is one of the applications running on Internet but it is the most significant of all.

By 1991, internat is made available to the common populace. Since then it has remained so.

2.1.2.5. The Mobile Revolution.

Analog Shrinkage doesnot proceed as linearly as Digital shrinkage but gain is remarakable. With digital you get more (more processing power and more functionality) for less (i.e.by scaling down). With Analog you get better performance wise by scalng down. 1Kg mobile phone slims down to your palm top. Base staions melted in size. Expensive infrastructure of telecom was turned into low cost consumer gear. Semiconductor miniaturization affected the entire Tel-com sector. Today in 2010 Tablet pc is becoming an universal communication appliance for getting information, for working, for learning, for transactions and for playing. How exactly this has happened.

In Figure 2.15, the typical Telecom Infrastructure is given.

Till 70s, all was analog.

Early 80s, digitization of Network Part.

Late 80s, Optical Fibers cable wasa installed on Trunk Lines-this is fiberization.

Early 90s, Packet switching was adopted as against circuit switching.

Late 90s, digitization of last mile.

Early 2000s, digitization of Terminal.

Late 2000s, wireless digiutization everywhere.

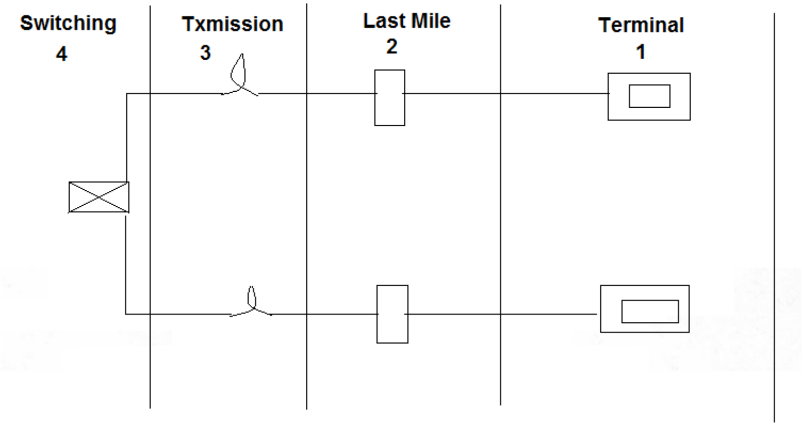


Figure 2.15. Telecom Infrastructure.

The consequences of Moore’s Law has been:

The increase in packing density from generation to generation with scaled down dimensions has enabled the integration of scaled up systems with greater functionalities and integration of bigger memory capacities in the same chip area.

The emergence of heterogenous processor chips has been the enabling technology of Mobile Revolution just as emergence of superscalar microprocessor chips were the enabling technology of Personal Computer Revoluion.

What is heterogenous processore chips?[“Calling All Phones – From a single-chip mobile communication to smart phones of the future”, Greg Delagi and Baher Haroum, Spring 2012, Vol.4,No.2, pp.21-25]

Historically, Intel and AMD focused on making CPU-based chips (e.g. Core and Athlon) while Nvidia focused on making standalone graphics chips (GPU) for the PC and server markets. All the other players in this landscape have utilized some form of system-on-chip system integration (SoC) to meet the diverse needs of their respective markets. This SOC is heterogenous procesor chip.

Power efficient and Cost efficient custom Digital Signal Processing(DSP) were developed for vocoders and digital communicatrion.

General Purpose DSP processor to support the rapidly evolving modem standards. If computer network uses our analog telephone network them MODEM will be the interface.

DSP increasingly played a bigger role versus analog electronics to realize signal paths initially for MODEMS and later for RF Cellular MODEMS.

Using these technologies, mobile phone, mobile phone size has been reduced and multiple communication functions have been integrated into a single chip. This is called System-on-Chip (SOC). The generic definition of a SoC is the on-chip integration of a variety of functional hardware blocks to suit a specific product application.A SoC can thus be as simple as a basic connectivity chip which combines some mixed-signal and digital circuitry.A more complex SoC may include the on-chip integration of an application processor unit (APU) and a graphics processor unit (GPU).Even more functional SoCs further integrate various other hardware blocks (e.g. image processor, audio/video decoder and modem).It is this ability to continue to integrate disparate functionality on a chip that has enabled SoC technology to rapidly evolve from supporting a simple feature-phone to a smartphone and all the way to a tablet computer.Using low power wireless links and energy efficient computations, multi-media and internet access have been accommodated on a single chip. Multi-medis content is imaging, graphics, video, display and HDI. With exploding consumption of multi-media content, with the advent of 3D Movies and with new content creation capabilities, Smart Phones multi-media related demand on computation is increasing at an unprecedented pace. As size and resolution of the images and video increase exponentially the computational need of the smart phones will correspondingly increase.

Qualcomm started out by designing chips for the growing connectivity market with the advent of cellular telephony and the internet.Qualcomm evolved its product line-up from standalone connectivity chips by adding: an applications processor (Krait via ARM license), a GPU (Adreno via AMD Imageon buyout) and a power management unit.Qualcomm’s flagship products (Snapdragon family) now include all these blocks making it a highly functional mobile SoC product.

Qualcomm (San Diego) has reigned supreme in the market for application specific handset IC like baseband and RF chips since 2007, IHS said. Last year, Qualcomm captured 31 percent of all handset core IC revenue, according to IHS’s wireless competitive landscape report. Qualcomm’s market share was up from 23 percent in 2007, IHS, said.

In 2007, the South Korean giant did not even rank in the top 10 in handset IC sales, according to IHS. By last year(2012), Samsung held firm to the No. 2 spot, capturing 21 percent of all handset core IC revenue, according to the IHS report.

Paradigm Shift

The rise of smartphones and 4G LTE have created “paradigm shifts that transformed competition” in the market for mobile handset core ICs.IHS defines the cellphone core IC market as including chips that provide mobile handsets with wireless wide-area-networking (WWAN) communication and application-processing capabilities. The market segments include handset core ICs for analog baseband, digital baseband, power amplifiers, radio and intermediate frequencies, high-level operating systems and software processors, and other multimedia or graphics coprocessors.

ARM(Advanced RISC Machines) are heterogenous processor or SOC. It hosta an embedded processor for user interface, for web browsing and a general purpose DSP with highly – power efficient data path accelerator for multi-media executions. These heteogenous processor chips more commonly known as SOC enable a long lasting multi-media experience without drawing on mobile's limited battery. If this is not ensured smart phones would not gain in popularlity and would not become a common place device as a personal hub for present and future needs.

Today most sophisticated heat management techniques are being used so that at peak power consumption it remains within maximum temperature permitted. Improved energy per application function and lower leakage power have to be coupled with architecture innovation in IC Design. These architecture innovation can be multi-core or asymmetric multi-core. Te reduce the peak power in the mobile set a number of functions are outsourced to the cloud computing.

Ubiquitous Android in the mobile world

Google Android will be to the next decade what Microsoft Windows was to the 1990s.It will be the software platform that will enable many of the most interesting and diverse devices to merge in electronics.For decades, the industry has sought—usually in vain — a common base of free, open-source software.The 1980s saw the quest for a unified Unix for computers.More recently, the search was for a single version of mobile and embedded Linux to power everything from mainstream smartphones, tablets and connected TVs to systems on the factory floor.Google's Android has come closer than anything in the past to fulfilling the dream.As ARM(Advanced RISC{Reduced Instruction Set Computer} Machines)-based processors strengthen their processing punch,Android will emerge as an operating system for notebooks , PCs and mobile sets.In smartphones, Android has already surpassed Apple's iOS and other alternatives in shipments.Its broad support among handset makers is its best guarantee of a robust and long life for use in all systems.Google's recent work to merge tablet and smartphone variants in version 4.0, called Ice Cream Sandwich, has put the codebase on a solid footing, at least for the moment.AMD, Apple, Dell, Dialog, Freescale, Fujitsu, HP, LSI, Microsoft, Motorola, MStar, Nvidia, Qualcomm, Samsung, STMicroelectronics and Texas Instruments- these fabless companies are all licensing ARM cores for everything from smartphones to tablets to basestations to servers.The truth is that even though ARM is gaining market share in low-power embedded applications the company still has a long way to go to challenge the perennial microprocessor king-INTEL.Intel in particular has a 20-year head start over ARM, resulting in a maturity, sophistication and veneration that will be hard to displace by a 12-year-old ARM.

Bridging the Data Deluge Gap--The Role of Smart Silicon in Networks

The proliferation of smart mobile devices, video, user-generated content and social networking, and the rising adoption of cloud services for both enterprise and consumer services are all driving explosive growth of wireless networking infrastructure.Globally, mobile data traffic is expected to grow 18-fold between 2011 and 2016, reaching 10.8 exabytes per month by 2016. Today, video traffic alone accounts for 40 percent of the wireless network load.The number of mobile devices connected to wireless networks will reach 25 billion, averaging 3.5 devices for every person on the planet, by 2015.That number is expected to double, to 50 billion, by 2020. This growth in storage capacity and network traffic is far outstripping the infrastructure build-out required to support it, a phenomenon known as the data deluge gap.

The bridging of the Data Deluge Gap.

To bridge this gap, the industry needs to leverage smarter silicon technology to scale data center infrastructures more cost effectively.Besides helping close the data deluge gap, smarter data processing offers potential dramatic improvements in application performance.A recent survey of 412 European data center managers conducted by LSI revealed that while 93 percent acknowledged the critical importance of improving application performance, a full 75% do not feel that they are achieving the desired results. This indicates that there is rising pressure on data center managers to find smarter ways to push systems to do much more work within the same power and cost profiles.Smart software running on general-purpose processors, increasingly with multiple cores, is pervasive in the data center.General Purpose Processors have long inhabited switches and routers, firewalls and load-balancers, WAN accelerators and VPN gateways.None of these systems are fast enough, however, to keep pace with the data deluge on its own, for a basic reason: general-purpose processors must treat every byte equally. While such equality is perfectly acceptable for system-level versatility, it is inadequate for low-level, high-volume packet processing.

Accelerating Networks

In many organizations today, microseconds matter, driving strong demand for faster response times.For trading firms, latency can be measured in millions of dollars per millisecond.For others, such as online retailers, every millisecond of delay can mean lost sales and fading customer loyalty.Tomorrow’s data center networks will need to be both faster and flatter,and therefore, smarter than ever.To eliminate the data deluge gap and maximize performance,systems need to be smarter,and those smarts will increasingly need to take the form of purpose-built silicon.

Future Predictions

Baseband chips, already accounting for more than half the revenue of the total handset core IC space, will maintain their pre-eminence in determining the market-share gains and losses of industry vendors moving forward, IHS said.But the future will also be driven by the ability of any given IC supplier to provide platform solutions that optimize the system-level design of all of the ICs, making up the handset's core chip architecture, the firm predicted.

Advent of multi-touch intuitive interface and its dependent application has increaded the expectations from the User Interface. Today following are our expectations from our User Interface namely:

1. Touchless gestures;
2. Hand, hghead/eye position awareness;
3. Haptic feed-back;
4. Other environmental sensing capabilities via multiple sensor nodes.

These User Interface capabilities are already available in health monitoring, exercise and environmental control tools. These capabilities are also coming on wrist watches, smart clothing and smart shoes. In order that these subtle capabilities are realized we need smart phones to be “instantly on, always on, always connected and always aware.” This will enable:

1. Accurate and continuous location monitoring both indoor as well as outdoor.
2. Smart. low power sensor and actuators network,
3. Ultra low power agents.

The advent of ultra low power wireless sensor nodes in the immediate environment of smart phones will provide valuable local information that requires constant awareness from the smart phone. Smart phones will need to interact with resources in the environment- this will require high data rate links with high energy efficiency of transmission. The huge increase in data communication requires significant reduction in the amount of energy spent per bit when transmitting all the data.

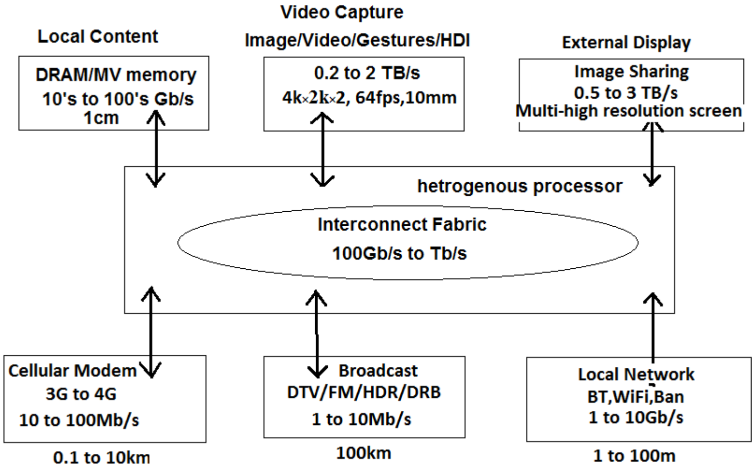


Figure 2.16. The toxonomy of Data and Rate Progression in a smart phone.

The moble revolution has been enabled by Silicon Technology Scaling but this is not adequate to meet the exponentially increasing needs of computation. We need 100 times greater enery efficieny to meet the exploding computational needs. To this end we need the following:

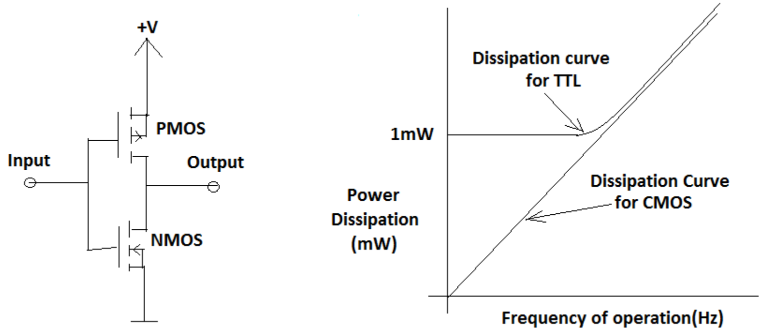
1. New architecture for complex computations;
2. New communication mechanism for short and long distances to break the current limit on the energy per bit communicated.
3. New method for all round environment sensing.

“Collectively the resulting advances will provide wealth of information to the key-hub and the ultimate link to the cloud – the smart phone – catapulting us into an exciting new technological world.”

2.1.2.6. CMOS Technology is the technology of choice for all applications as we make a transition from micro to nano era.

[“The Age of Digital Nomad- Impact of CMOS innovation’, Tsugio Makimeto, IEEE Solid State Circuits Magazine, Winter 2013, Vol.5, No.1]

A CMOS inverter is a nano-watt logic because its stand-by power dissipation is in nano-watt. Hence right from the beginning it has been the Technology of Choice in microprocessor design and fabrication. Its circuit configuration and its power dissipation curve are shown in Figure 2.17.


+V
PMOS
Input
Output
NMOS
Dissipation curve for TTL
1mW
Power Dissipation (mW)
Dissipation Curve for CMOS
Frequency of operation(Hz)

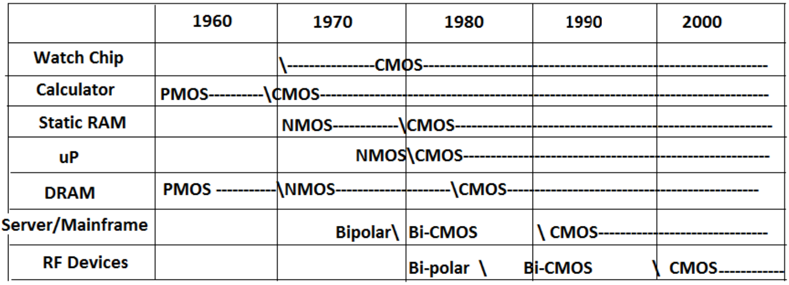
Input State	PMOS	NMOS	OUPUT	Output State
'0'	ON	OFF	Source	'1'
'1'	OFF	ON	Sink	'0'

Figure 2.17. Circuit Structure of CMOS and its dissipation curve in comparison to TTL.

As we move from Micron Technology (Channel length is greater than 1µm) to sub-micron Technology(Channel length is between 990nm-100nm) to deep-submicron

Technology(Channel length is between 100nm-50nm) to ultra-deep-submicron Technology(Channel length is between 50nm-10nm), power management and power dissipation are becoming bigger and more crucial issue hence all applications are converging to CMOS technology because of its zero stand-by feature as seen in Figure 2.18.

There was clear demarcation among consumer, computer and communication systems but with the advances of Semiconductor Technology they are collectively called digital consumer products. This market convergence and this market convergence is being driven by digitization of information and CMOS innovation.



	1960	1970	1980	1990	2000
Watch Chip		\---------------CMO	S----------------------	----------------------	----------------------
Calculator	PMOS----------\	CMOS----------------	----------------------	----------------------	----------------------
Static RAM		NMOS-----------\	CMOS------------------	----------------------	----------------------
uP		NMOS	\CMOS-----------------	----------------------	----------------------
DRAM	PMOS -----------	\NMOS---------------	--------\CMOS---------	----------------------	----------------------
Server/Mainframe		Bipolar\	Bi-CMOS	\ CMOS----------------	----------------------
RF Devices			Bi-polar \	Bi-CMOS	\ CMOS-----------

Figure 2.18. All applications including RF applications are convergfing to CMOS Technology.

Qualcomm pits CMOS against GaAs for cellphone front ends

Only recently QUALCOMM has introduced CMOS Power Amplifiers in the front end of Mobile Phones."Qualcomm is the first to launch a CMOS PA [power amplifier] alternative to GaAs-based multiband multimode PAs for mid- to high-tier

3G/4G smartphones, a phone segment formerly the exclusive domain of GaAs PAs.Envelope tracking power management enables this, and we expect adoption of CMOS PAs from Qualcomm and others to accelerate in mobile phones as a result," said Christopher Taylor, director of an RF and wireless component market research service at Strategy Analytics.CMOS PAs will give a cost advantage.Strategy Analytics say "the new PowerAmplifiers and EnvelopeTracking(ET) from Qualcomm’s GaAs-oriented competitors will dilute Qualcomm’s new presence in the RF front end, but widely available ET will also improve prospects for CMOS PAs, challenging the dominance of leading GaAs-based PA suppliers Skyworks, RFMD and others".

Today the hall mark of Solid State Systems are still more clearly defined and are visible:

1. More intelligence(or higher information processing capability);
2. Smaller size- better portability;
3. Lower power consumption for longer battery life;
4. Lower cost (for more affordability as personal devices).

Figure of Merit = Intelligence/(Size*Cost*Power). Here intelligence is defined as MIPS(Mega Instruction per second) for general purpose computers; MOPS(Mega Operations per second) for signal processor and FLOPS(Floating Point Operations per second) for numerical calculating systems.

Semiconductor Technology has made great progress towards maximizing the Figure of Merit.

SSPD_Chapter 2.1.3.Nano-Technology Era.
SSPD_Chapter 2.1.3.The present Nano-Technology Era in which Semiconductor Industries have formally entered with the discovery of exfoliated Graphene in October 2004.

2.1.3.Nano Technology Era (October 2004 onward)

[A new class of materials Two-dimensional(2D)crystalline materials have recently been identified and analyzed (1). The first material in this new class is graphene, a single atomic layer of carbon. This new materialhas a number of unique properties, which makesit interesting for both fundamental studies and future applications.The electronic properties of this 2D-material leads to, for instance, an unusual quantum Hall effect.(2,3)It is a transparentconductor(4) which is one atom thin. It also gives rise to analogies with particle physics, including an exotic type of tunneling(5,6) which was predicted by the Swedish physicist Oscar Klein.(7).In addition graphene has a number of remarkable mechanical and electrical properties.

It is substantially stronger than steel, and it is very stretchable. The thermal and electrical

conductivity is very high and it can be used as a flexible conductor. The Nobel Prize in Physics 2010 honours two scientists, who have made the decisive contributions to this development.

They are Andre K. Geim and Konstantin S. Novoselov, both at the University of Manchester, UK. They have succeeded in producing, isolating, identifying and characterizing graphene.(1,8).

Excrpted from: http://www.nobelprize.org/nobel_prizes/physics/laureates/2010/advanced-physicsprize2010.pdf.

References:1. K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, Science 306, 666 (2004).

2. K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, I. V. Grigorieva, S. V. Dubonos, and A. A. Firsov, Nature 438, 197 (2005).

3. Y. B. Zhang, Y. W. Tan, H. L. Stormer, and P. Kim, Nature 438, 201 (2005).

4. R. Nair, P. Blake, A. Grigorenko, K. Novoselov, T. Booth, T. Stauber, N. Peres,

and A. Geim, Science 320, 1308 (2008).

5. M. I. Katsnelson, K. S. Novoselov, and A. K. Geim, Nature Physics 2, 620 (2006).

6. A. F. Young and P. Kim, Nature Physics 5, 222 (2009).

7. O. Klein, Z PHYS 53, 157 (1929).

8. K. S. Novoselov, D. Jiang, F. Schedin, T. J. Booth, V. V. Khotkevich, S. V. Morozov,

and A. K. Geim, Proceedings of the National Academy of Sciences of the United States of America 102, 10451 (2005).]

We have come a long way since the first IC chip was invented in 1959 in TI and FairChild simultaneously. The invention of IC Chip has marked the beginning of a new historical era in Man-Kind’s history. This invention has been as crucial as the discovery of fire was or as momentous as the invention of wheel was.

Discovery of Fire brought the proto-human society down from its arboreal abode to its humble cave dwellings. This enabled the development of our society in rapid strides.

On the other hand the invention of wheel led to rapid exchange of ideas and commodities enabling a homogenization across the globe which would have been impossible otherwise.

In a similar fashion IC Technology has been a great leveler of human societies. All human beings are becoming a knowledge worker in one way or the other and the whole world is being rapidly transformed into a big Global Village where all new ideas or new technologies are available to one and all in the shortest time.

Moore's Law prediction of increase in packing density from generation to generation with scaled down dimensions has enabled the integration of scaled up systems with greater functionalities and integration of bigger memory capacities in the same chip area.

Table 2..1.3.1. Clasification according to the scale of integration.

SSI	MSI	LSI	VLSI	ULSI
<30comp.	30 To 100comp.	100,000comp.	>1million comp.	>100million comp.
RTL	Mag.Comp.	4004μP Chip	Pentium IV	Corei7

Table 2.1.3.2. MOS Dimension Scaling.

MOS	1967	1997	1999	2001	2003	2006	2013
L(μm)	10	0.25	0.18	0.13	0.10	0.07	32
DRAM(Gb/cm2)	64M	0.18	0.38	0.42	0.91	1.85	32
Junction Depth(xj)nm	1000	100	70	60	52	40	?
Interconnection Pitch(nm)	2000	600	500	350	245	70	32
Technology	micron	Sub-micron technology				Deep sub-micron	Ultra-Deep-Sub-micron

Table 2.1.3.3.Processor scaling from 2006 to 2012.

Processor	MOS count	Year	Manf.	Process	Area(mm2)
AMD K10 quad core2ML3	463M	2007	AMD	65nm	283
AMD K10 quad core6ML3	758M	2008	AMD	45nm	258

Corei7(quad)	731M	2008	Intel	45nm	263
6Core Opteron 2400	904M	2009	AMD	45nm	346
16Core Sparc T3	1B	2010	SUN/ ORACLE	40nm	377
6Core Corei7	1.17B	2011	INTEL	32nm	216
QuadCore + GPU Corei7	1.4B	2012	INTEL	22nm	160
62Core Xeon Phi	5B	2012	INTEL	22nm	?

Table 2.1.3.4.GPU scaling from 1997 to 2012. [GPU – Graphical Processing Unit is a specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the building of images in a frame buffer intended for output to display]

Processor	MOS count	Year	Manf.	Process	Area(mm^2)
NV3	3.5M	1997	NVIDIA	350nm	90
Tahiti	4.3B	2011	AMD	28nm	365

RV1070					
GK110Kepler	7.08	2012	NVIDIA	28nm	561

Table 2.1.3.5.FPGA (Field Programmable Gate Array). It can be configured by the user according to his sysyem requirements. With increase in components the options for configurability is getting enhanced day by day.

FPGA	**MOS count**	**Year**	**Manufacturer**	**Process**
VIRTEX	70M	1997	Xilinx	?
VITEX II	350M	2000	Xilinx	130nm
Virtex 4	1B	2004	Xilinx	90nm
Stratia IV	2.5B	2008	Altera	40nm
Virtex 7	6.8B	2011	Xilinx	28nm

We have already seen in Section 2.1.2.6. that CMOS, because of its nanowatt power dissipation in stand-by mode, is becoming the technology of choice at the lecvel of ULSI. All applications are converging to CMOS process technology.

Theoretical proposition of CMOS as nano-watt logic family was made in1963 at International Solid-State-Circuits Conference(ISSCC-1963).

In 1968, RCA commercialized CMOS series of logic family known as CD4000 series. This was equivalent to TTL 74series.

Till 1970 the Industry Consensus was:

1.Mainstream device will remain NMOS.

2.BJT will be for Analog applications and for very high speed applications.

3.PMOS will be phased out.

4.CMOS will remain the process technology for Wrist Watches ICs.

This consensus was driven by the economic criteria. CMOS technology had a high cost factor. But in 1978, Hitachi pioneered a more economic way of fabricating CMOS products based on the invention of twin well CMOS by Y.Sakai and T.Masuchari then at Hitachi Center for Research Lab. In 1978, Hitachi marketed 4kSRAM based onCMOS which outperformed Intel's equivalent 4kSRAM based on NMOS technology.

Table.2.1.3.6.Comparative study of 2147/ INTEL(1977) and 6147/Hitachi(1978).

Product	HMOS 4kStaticRAM	Hi-CMOS4kSRAM
Technology	NMOS	Twin well CMOS
Speed	55/70ns	55/70ns
IACTIVE/IIDLE	110mA/15mA	15mA/1μA
Chip Size	16.2nm2	11.5nm2

The Table clearly states that Hitachi CMOS memory is several orders of magnitude better than INTEL's NMOS memory as far as standby power dissipation is considered. By 1981, CMOS 16k SRAM was manufactured and sold.

Table 2.1.3.7 gives the comparative study of microprocessor(μP) based on NMOS and CMOS.

Table 2.1.3.7.Comparative study of 6801/HIT(79) and 6301/HIT(81)

Product	6801/HIT(79)8-bit μP	6301/HIT(81)8-bit μP
Technology	4 μmNMOS	3 μmCMOS
Speed	1MHz	1MHz,1.5MHz,2MHz
Power Active	900mW	30mW(1MHz)
Stand by Power	70mW	0.01mW
Pin Count	40Pins	40Pins

EPSON introduced all CMOS computer HX-20 in 1982. It was a hand-held computer and proto-type/ ancestor of the present day nomadic tools. Gradually all products are shifting to CMOS Technology.

Presently Digital Consumer Market(smart phones and Tablet PCs) is driving Semiconductor Industries. In coming day, Automotive market, Robotics, Medfical& Health Care and Sensor Network Market will be driving the Semiconductor Market.

The processor speed has increased 10 times from Pentium in 1993 to Pentium III in 1999 but with the introduction of Core2Duo in 2006 to Corei7 in 2013 the core speed has increased 1.5 times only but the appetite for high speed computing is insatiable. So obviously this paradigm of core computing will not be able to keep up with the computing needs of the post-industrial era. So new computing paradigms will have to be invented.

Breahthrough in Strained Silcon and High-K,metal gate Technology has allowed Fabless Companies to follow the Road Map set by Moore’s Law.

2.1.3.1. 3-D Tri-Gate Technology/Fin-FET Technology.

In 1991 a revolutionary 3D FinFET concept was put forward by Digh Hisamato and team of other researchers at Hitachi Central Research Laboratory.

This proposed “WRAP AROUND” gate transistor technology. In conventional Planar Technology, GATE exercises control on the 2-D channel of electrons/holes only from the TOP.This new technology allowed the gate to exercise its control from 3 directions: From the TOP as well as from the TWO SIDES as shown in the Figure 2.19.

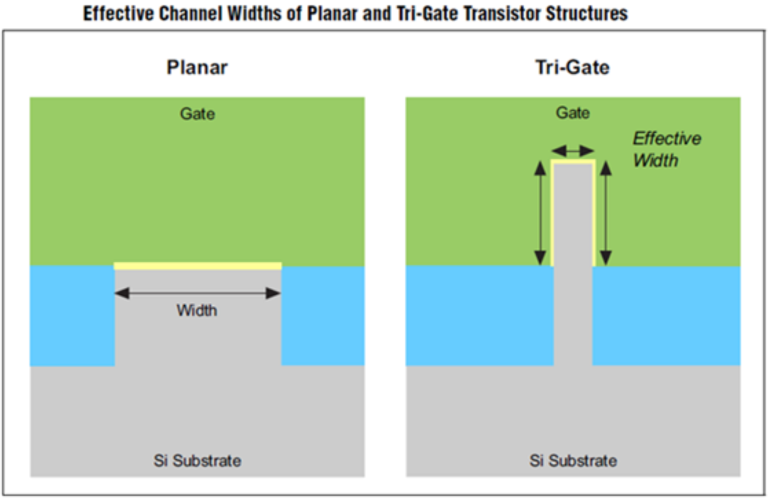

Effective Channel Widths of Planar and Tri-Gate Transistor Structures
Planar
Gate
Width
Si Substrate
Tri-Gate
Gate
Effective Width
Si Substrate

Figure 2.19. Planar Gate Control and Tri-Gate Configuration in Fin-FET.

In Figure 2.19, the channel widths of Planar Transistor and Fin-FET are defined. In Figure 2.20 the channel length is defined in Fin-FET. The current drive capability and performance is directly proportional to its effective channel width as seen from the four Equations below:

$$Triode-Region\_I_D = \frac{W}{L}\mu_n \frac{\varepsilon_{ox}}{t_{ox}}(V_{GS} - V_{Th})V_{DS};\_\_\_\_\_\_(2.1.3.1)$$

$$Pentode-Region\_I_{DSat}=\frac{W}{2L}\mu_n\frac{\varepsilon_{ox}}{t_{ox}}(V_{GS}-V_{Th})^2;\_\_\_\_\_\_\_(2.1.3.2)$$

$$Triode-Region\_g_m=\frac{W}{L}\mu_n\frac{\varepsilon_{ox}}{t_{ox}}V_{DS};\_\_\_\_\_\_\_(2.1.3.3)$$

$$Pentode-Region\_g_m=\frac{W}{L}\mu_n\frac{\varepsilon_{ox}}{t_{ox}}(V_{GS}-V_{Th});\_\_\_\_\_\_\_(2.1.3.4)$$

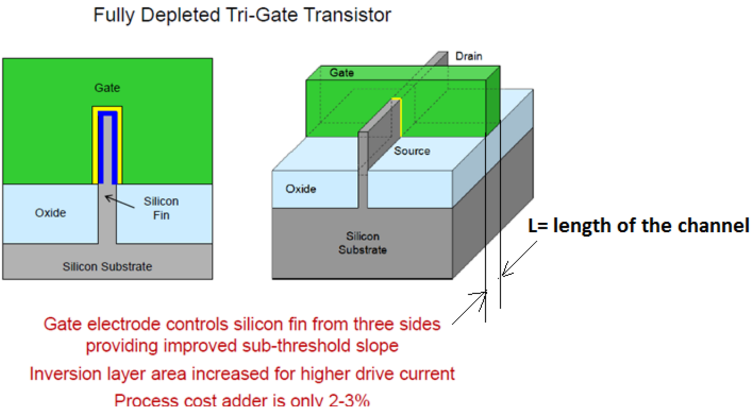

Fully Depleted Tri-Gate Transistor
Gate
Oxide
Silicon Fin
Silicon Substrate
Drain
Gate
Source
Oxide
Silicon Substrate
L= length of the channel
Gate electrode controls silicon fin from three sides providing improved sub-threshold slope
Inversion layer area increased for higher drive current
Process cost adder is only 2-3%

Figure 2.20. Definition of Length of the Gate.

The channel width can be significantly enhanced in 3-D Transistor structure in relation to planar structure because of the ability to extend the the width in the third dimension without increasing the layout area. This provides the potential for both enhanced design flecxibility for the designer as well as increased performance withput the same penality in 2-D area which exists when enhancing the width in planar transistor.

There is reduced active and leakage power in 3D Fin-FET transistors. This advantage results from the improved control of the channel by Gate’s electric field from the three sides of the Fin as shown in the figures. This results in a steeper curve of I_D-V_{GS} for Fin-FET as shown in Figure 2.21.The steeper curve results in significant reduction in leakage current in Fin-FET for the same performance of planar transistor or substantially higher performance in terms of operating speed or a combination of both.

The power supply voltage can be significantly reduced with Tri-Gate without impairing its speed of operation due to increased effective width in Tri-Gate compared to planar transistor. The combination of lower supply voltage and reduced leakage current results in substantial power saving.

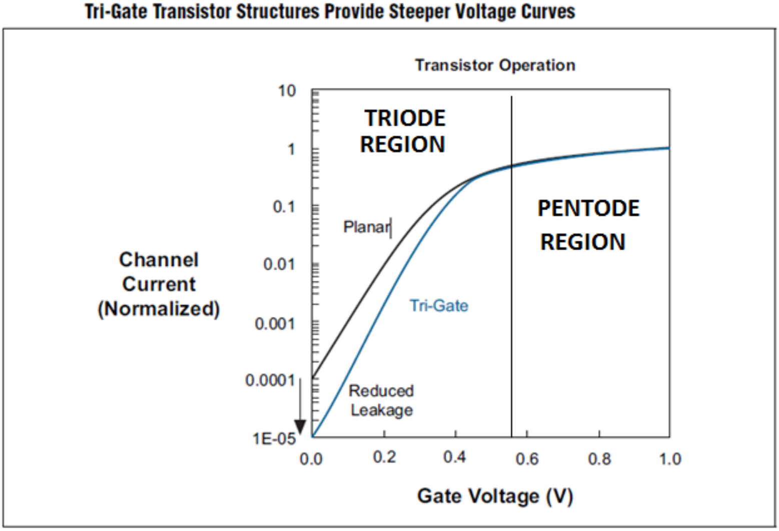

Tri-Gate Transistor Structures Provide Steeper Voltage Curves

Figure 2.21. Id vs Vgs in planar and Fin-FET Transistors

Fin-FET has a smaller cross-sectional area for the same performance of planar transistor.This gives a reduced area of incidence for cosmic particles which result into reduced Single-Event Error(SEE). This has been authenticated by 22nm implementation of Intel TriGate Fin-FET.

3-D Geometry and structure provides a host of improvements over planar structure as discussed above and enumerated below:

1. Improved performance;
2. Reduced active and leakage power;
3. Higher packing density;
4. Reduction in susceptibility to charged particles SEE.

In 1974, IBM's Robert Dennnard, inventor of single transistor DRAM(Dynamic RAM) showed that when MOSFET geometries, voltages and dopings were scaled down, gate transit time also scaled down and performance thus improved by the same factor. ["Design of Ion-Implanted MOSFET with very small Physical Dimensions", *IEEE Journal of Solid State Circuits,* Vol.SC-9,No.5, 256-268,October 1974]. Taken together, density improvement as predicted by Moore's Law and the performance improvement by Dennard signaled a coming explosion of growth in chip processing power. This precisely happened and this is what has been fueling Computer-Communication Revolution till date.

But there was great setback in 2005. Traditionally each new generation of Simanufacturing involves higher density and higher speed. But this scalng trend predicted above started plataeuing out at 90nm process technology. Clock rates stagnated as already pointed out above in Section 2.1.3. Power consumption and dissipation have grown with scaling. There was unacceptable gate leakage current. Therefore high-K(HfO, $K = 25$) and metal gate technology was adopted. This maintained the improvement of performance by scaling till 40nm process technology.

On 4th May 2011, Intel adopted Fin-FET (3D) Transistor for the manufacture of 22nm Semiconductor Products. Intel's 22nm Tri-Gate uses 3rd Generation high K/metal-gate scheme, copper interconnect, strained Si and 193nm Immersion Lithography.

International Technology Roadmaps for Semiconductor (ITRS) in 2012 identified 3-D Transistor Technology as the primary enabler of all incremental semiconductor improvements beyond 22nm process technology. 3-D Tri Gate is able to maintain the advantage below 22nm of higher packing density not by scaling but by extending the transistor width in the third dimension. This allows designers the ability to trade off the size and width of the transistor 'fin' based on performance, power, and transistor density packing objectives.

By adopting the 3-D technology, the steady march and consistent improvement in packaging density according to Moore’s Law has been maintained. This has meant steady improvement in performance, reduction in power back-up and reduction in cost per transistor. In addition there has been steady improvement in standby power and operating power dissipation.

2.1.3.2. Tri-Gate Devices now in production.

While the advantages of Tri-Gate transistors have been studied and known for some time, adoption and implementation is driven ultimately by technology and

manufacturability, as well as cost-effectiveness.

The advanced state of semiconductor manufacturing at very small geometries (40 nm, 28 nm, 22 nm or 20 nm and beyond) requires research and development expenditures

that now limit this technology to a handful of companies with capital expenditure capabilities in the billions of dollars. As a result, only a handful of manufacturers are

able to capitalize on the known advantages of 3-D transistor technology. Intel Corporation is the only company to have made this design and manufacturing transition in 22 nm technology, and can provide data on the overall maturity and manufacturability of Tri-Gate transistors on a mass production scale. This data, as of the first quarter of 2013, includes 100 million units of Tri-Gate transistor-based products.

Several known issues and characteristics of the 3-D gate structure have been acknowledged and addressed to achieve manufacturing and design maturity with the technology. These include the modeling of new parasitic capacitance values not modeled in traditional planar designs, layout dependent effects, and the use of double-patterning techniques using current lithographic equipment to form closely spaced fins.

The electronic design automation (EDA) community is also an important factor in the maturity and usability of FinFET and Tri-Gate design technology to the semiconductor designer. A great deal of publicity and user education is underway in

2013 by companies like Cadence and Synopsys revolving around the impact of

Tri-Gate rules and flexibility in the design of future semiconductor products.

Intel has maintained its leadership in innovation in technology process so as to maintain the march along the Road Map laid down by Moore’s Law. Figure 2.22 shows the technology innovation introduced by INTEL in last decade of Semiconductor Manufacturing history.

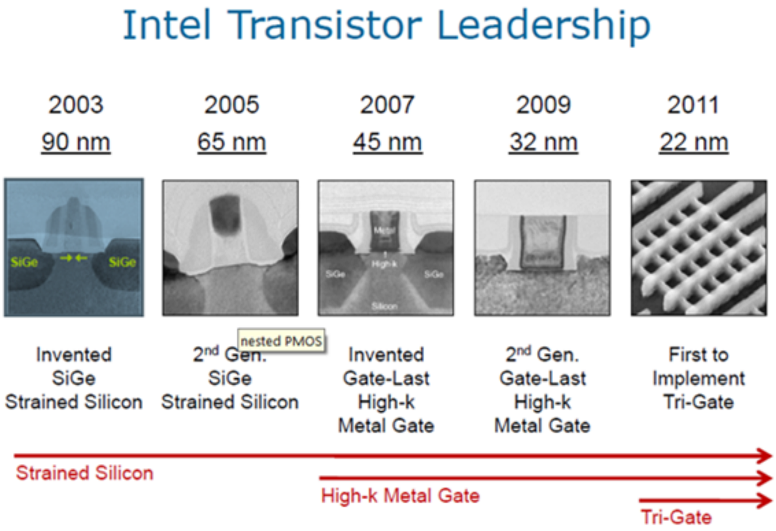

Intel Transistor Leadership
2003
90 nm
2005
65 nm
2007
45 nm
2009
32 nm
2011
22 nm
SiGe
SiGe
Metal
High-k
SiGe
SiGe
Silicon
nested PMOS
Invented SiGe Strained Silicon
2nd Gen. SiGe Strained Silicon
Invented Gate-Last High-k Metal Gate
2nd Gen. Gate-Last High-k Metal Gate
First to Implement Tri-Gate
Strained Silicon
High-k Metal Gate
Tri-Gate

Figure 2.22. The innovation made in process technology by Intel so as to maintain the march along the Road-Map set by Moore's Law.

In February 2013, Altera and Intel Corporation jointly announced that the next generation of Altera’s highest performance FPGA products would be produced using

Intel’s 14 nm 3-D Tri-Gate transistor technology exclusively. This makes Altera the exclusive major FPGA provider of the most advanced, highest performance semiconductor technology available.

2.1.3.3. Graphene Technology-the future generation IC enabler.

All this advancement and convergence in computation and communication was made possible due to scaling. But this is becoming more difficult as we go below 20nm process.[Hennessy, J; Patterson, De; “Computer Architecture”, 3rd Edition, Morgan Kaufman Publishers, Amsterdam 2003].

According to International Technology Roadmap for Semiconductors (ITRS) we have the projection as given in Table 2.1.3.8

Table 2.1.3.8. Roadmap of CMOS Technology according to ITRS-2003

Year	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
LGate (nm)	32	28	25	22	20	18	16	14	13	11	10	9
	←Bulk/ PartiallyDepleted-SOI→											
				←Fully Depleted – SOI→								
							←Multi-gate→					

SRAM uses 6-Transistor to store a bit. SRAM is dominant form of embedded memory. It comprises 60- 70% chip area and 75% - 85% transistor count. As the logic gates density is increasing correspondingly the memory density must increase.

Statistical dopent fluctuations, variation in oxide thickness and line edge roughness increases the spread in VTh and ION/IOFF in nano-scale regime [Murari 2003]. Increased leakage and parametric variations are making scaling of 6-T SRAM memory array difficult.

Due to fundamental physical limit to scaling, the era of conventional linear scaling of transistor dimensions has ended. Power dissipation and process induced variations have become the major issues for continued scaling of Si MOSFETs in future generation.

New materials and device structure are needed to continue the scaling trend as seen in ITRS-2003 Road map given in Figure 2.23.

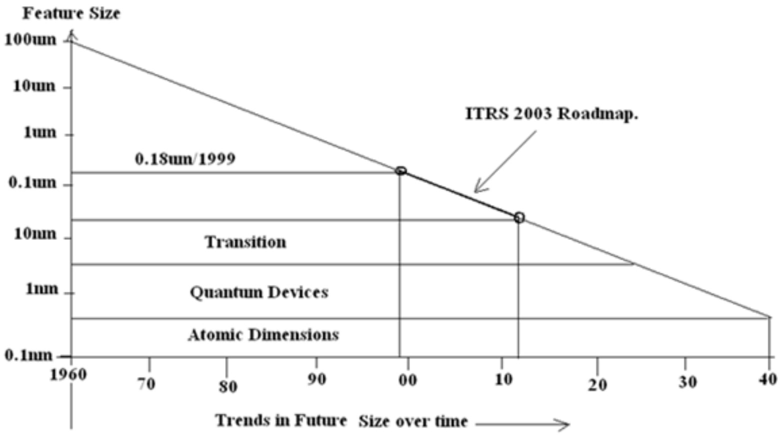


Figure 2.23. Trends in Future Size as illustrated in ITRS-2003 Roadmap.

According to ITRS 2003, by 2014 L_{Gate}= 35nm, V_{DD} = 0.4V and clock frequency = 30GHz.

Shrinking geometries, low power voltages and high frequency have a negative impact on reliability. These enhanced features have increased the failure rates.

2.1.3.3.1. New and innovative methods to meet the challenges of nano miniaturization.

Multiple core with threading has been one innovation for solving the thermal management at nano level.

Copper interconnect provides higher electromagnetic threshold than Aluminum. So Industry is adopting Cu interconnects.

Fault avoidance and fault tolerance are the possible ways of improved reliability.

Fault avoidance relies on improved materials, improved manufacturing processes and improved circuit design. Cu interconnects is an example of improved materials.

SOI (Silicon-on-Insulator) is the process solution for lower circuit sensitivity to particle induced transients. Fault tolerance is implemented at circuit and system level. It relies on error detection and error correction code.

1. **Future Generation Integrated Circuits.**

Human appetite for computation has grown faster than the processors power of real life computers. We need more powerful processors just to keep up with modern applications like interactive multi-media, mobile computing and wireless communication. In addition battery power is at premium in mobile computing environment hence power aware components/systems are indispensible.

Low power design can be achieved at different level of the system- basic process/device level (threshold reduction, multi-threshold deices), at circuit level (transistor sizing, logic optimization, activity driven power down, low swing logic, adiabatic switching), at architecture level (voltage scaling, parallelism,, instruction set , signal correlation) at algorithm level(complexity, concurrency, locality, regularity, data representation) and at system level.(design partitioning and power down).

We have to simultaneously look at performance-power-cost product . This will lead to new computer architecture.

As we scale down the dimensions for each new generation technology, new areas of applications are emerging.

To continue its historical growth and continue to follow Moore's Law, the Semiconductor Industry will require advances at all fronts- from front end processors and lithography to design innovations, high performance process architecture and SOC solutions.

Further down the scaling strategy we are bound to hit a road block.

The recent development in the field of Graphene Transistors show that these could be a promising candidate for taking over the mantle for continued scaling down according to Moore's Law.[Kreupl, Franz " Carbon nano-tubes finally deliver", *Nature,*.**484**, 321-322, 19th April 2012; Franklin, Aaron D.;Luisier, Mathieu; Han, Shu-Jen;et.al. “Sub-10nm Carbon Nano tube Transistors”, *Nano Letters Nature,* **12,** 758-762 (2012)].

The present Graphene Transistor using single walled Carbon nano-tube(SWCNT) of 9nm at a supply voltage of 0.4V switches from ‘0’ to ‘1’. The current ratio of ON to OFF is 10,000 times which is excellent for LOGIC applications.. This operates at low voltage hence it mitigates the demand in energy. The only difficulty is in fabricating these devices because of non-availability of SWCNT of one flavor.

In contrast according to IRTS 2011[ITRS2011 Edition , Front End Processors, Go.nature.com/ qykuh1(2011)], a similar sized Si MOSFET of 9nm gate length and operating at 0.64V will arrive in the market in 2022 and we cannot anticipate the cost.

So it is ripe time that we explore for alternatives if the present rate of miniaturization and development has to be maintained. The Future Generation Integrated Circuits is discussed in Chapter 8 of this Collectrion.

Supplement_2.1.3.2.Transport in NanoTechnology Era.
This supplement of Nano Technology Era gives the transport physics of charge carriers in nano-devices.

Supplement_2.1.3.2.Transport in NanoTechnology Era.

In micron size devices electron transport is that of a Fermion. Thermal energy causes random motion of electrons. It moves over a short distance and gets scattered thereby changing its direction. The magnitude of the mean free path decides its mobility. Mean Free Path in turn is decided by the crystal structure. If it is a single crystal with no defects, no impurities, no lattice vibration and no boundaries then the single crystal is a super-conductor with infinite mobility. Ordinarily the real life single crystals are non-ideal and hence they have finite mobility but strongly influenced by the boundaries, impurities, defects and lattice vibrations. In these micron size crystals, electron experiences diffusion from higher concentration to lower concentration as well as drift down the potential gradient.

As we reach sub-micron size devices and dimensions become comparable to the mean free path then electron experiences ballistic transport.

In Nano-meter size dimensions and in graphene, electron does not remain a Schrodinger Particle but becomes Dirac-quasi particle. It cannot be described by Schrodinger Equations but is described by Dirac Equations. Electron experiences Quantum Transport.

At 1nm and less it is atomic dimension and electron is in Quantum Transport region. These different modes of transport is shown in Figure A.

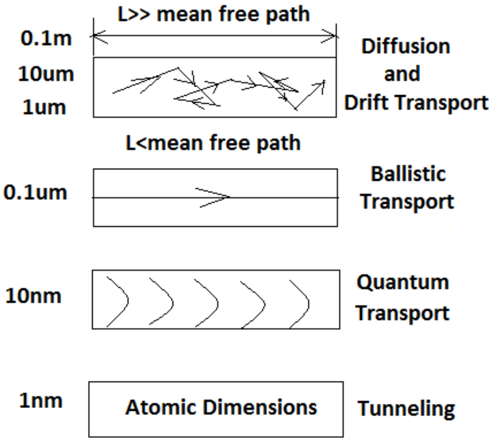
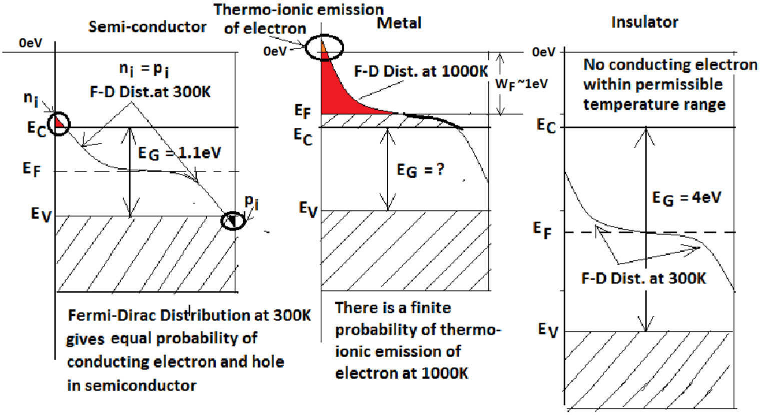


Figure A. As the dimension is reduced the nature of electron transport changes.

SSPD_Chapter2.2.Semiconductor Physics with emphasis on Silicon
SSPD_Chapter2.2 briefly describes intrinsic and extrinsic semiconductor and the behaviour of conducting electron and hole in semi-conductor with special reference to Silicon.

SSPD_2.2.Semiconductor Physics with emphasis on Silicon.

In Solid-State of Matter we have Insulators, Semiconductors and Metal. By definition Insulator is a bad conductor of electricity whereas Metal is a good conductor of electricity. Semi-conductor is intermediate conductor of electricity at Room Temperature 300K. Below 77K semi-conductor freezes out and becomes insulator. Therefore below 77K, Semi-conductor Devices stop working. This is a bad feature and a severe drawback. Such a freeze out does not occur in Graphene Devices.



Fermi-Dirac Distribution at 300K gives equal probability of conducting electron and hole in semiconductor

There is a finite probability of thermo-ionic emission of electron at 1000K

Figure 2.2.1.Comparative study of Semi-conductor,Metal and Insulator based on Energy-Band Diagram

E_C = lower edge of the Conduction Band.

E_V = Upper edge of the Valence Band.

Valence Band is completely filled up. Conduction band is completely empty at 0K in both Semi-conductor and Insulator. Conduction Band in Insulator remains empty within the permissible range of temperature.[Permissible range of temperature is 0K to melting point of the temperature]. But in Semi-conductor, above liquid Nitrogen Temperature (77K),the thermal agitation manages to break a few co-valent bonds and Electron-Hole Pair (EHP) are produced which give a small conductivity to semi-conductor. As temperature rises EHP rises exponentially given by the following relationship:

$$n_i^2 = N_C N_V Exp[-\frac{E_G}{kT}] \underline{\qquad\qquad\qquad} 2.2.1$$

$$n_i = \sqrt{N_C N_V} Exp[-\frac{E_G}{2kT}] \underline{\qquad\qquad\qquad} 2.2.2$$

N_C (=Effective Density of states at lower edge of the Conduction Band) ~ N_V(=Effective Density of states at upper edge of the Valence Band) = 10^{19} electrons or holes per c.c.

Therefore :

Intrinsic Concentration in Silicon (Band-

Gap=1.12eV) = n_i = p_i(at 300K) = 10^{10}/cc.

Intrinsic Concentration in Germanium(Band-Gap=0.67eV) = n_i = p_i(at 300K) = 10^{13}/cc.

Intrinsic Concentration in GaAs (Band-Gap =1.4 eV) = n_i = p_i(at 300K) = 10^{6}/cc.

In Figure 2.2.1. the comparative study of Semi-conductor, Metal and Insulator based on energy band-diagram is given. The Fermi-level (E_F) lies in the middle of the band-gap of both Semi-conductor and Insulator. In Metal the Fermi-level lies within the Conduction Band.

In order to fully appreciate the thermal generation of EHP we must briefly talk about Fermi-Dirac Distribution or Fermi-Dirac Staistics.

In this Universe electrons, protons and neutrons are real particles, have half integer spin angular momentum ($\pm 1/2\hbar, \pm 3/2\hbar\ldots$), occupy space and follow Fermi-Dirac Statistics.These are Fermions.

Photons, Phonons, Mesons, Gluons are quasi-particles, relativistic in nature, have integer spin angular momentum($0\hbar, \pm 1\hbar, \pm 2\hbar\ldots\ldots$) and a large number of Photons can pile up in a small volume.It follows Bose-Einstein Statistics.These are Bosons.

Here we will dwell upon Fermi-Dirac Distribution of Fermions.

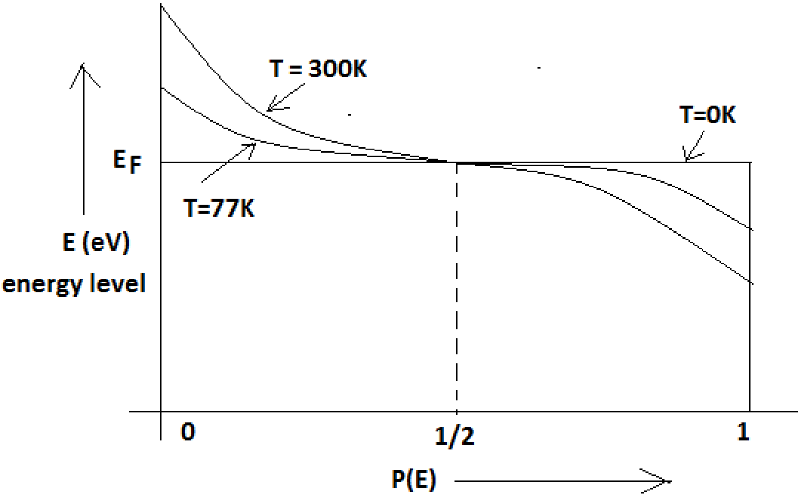


Figure 2.2.2. Fermi-Dirac Distribution of Fermions.

Mathematical formulation of Fermi-Dirac Distribution is the following:

$$P(E)=\frac{1}{1+Exp(\frac{E-E_F}{kT})} \text{\_\_\_\_\_\_\_\_\_\_\_\_} 2.2.3$$

Here P(E) = probability of occupancy of Energy Level E(eV).

T = temperature of the specimen in Kelvin and k = Boltzmann Constant.

E_F = Fermi-Energy Level. At this energy level, P(E) is ½ at a finite temperature.

At 0K (as is evident from Figure 2.2.3), the distribution is rectangular. What this means is that at 0K, all anergy levels upto Fermi-Energy Level is completely occupied where permissible and rest are completely empty.

As temperature is raised to 77K, distribution gets skewed. As temperature increases to 300K(Room Temperature)the distribution gets still more skewed as shown in Figure 2.2.2.

This is precisely why in Figure 2.2.1, the Fermi-Dirac Distribution is responsible for a low level of EHP generation by thermal energy in pure Semiconductor at Room temperature. As calculated from Equation 2.2.2, at Room Temperature GaAs,Si and Ge have equal concentrations of electron and holes equal to 10^{6}, 10^{10} and 10^{13} per cc. GaAs has the least concentration because it has the largest band-gap of 1.4eV. Hence GaAs behaves as semi-insulator. Whereas Ge has the largest concentration of electron and holes because it has the narrowest band-gap of 0.67eV.

Pure semiconductor is known as Intrinsic Semi-conductor and concentration of electron= concentration of hole = intrinsic concentration n_i.

In Figure 2.2.1, we also notice that Insulator typically has a band-gap greater than 4eV hence within the permissible temperature range Fermi-Dirac Distribution never becomes skewed enough so as to cause EHP generation by thermal agitation. Hence Insulator remains a bad conductor at all temperatures within the permissible range.

For Metal, high enough temperature (1000K) does cause thermal-ionic emission of electrons in case of metallic cathode with 1eV Work-Function as shown in Figure 2.2.1. Therefore Cathode material in Vacuum Tubes is always Sr-Ba-CuO(Strontium-Barium Copper Oxide).

SSPD_Chapter2.2.1.and 2.2.1.Solid State Equivalent of Vacuum Devices and Crystalline structure of Silicon

SSPD_Chapter 2.2.1 and Chapter 2.2.2 give the Solid State Equivalent of Vacuum Devices and the Crystalline Structure of Silicon.

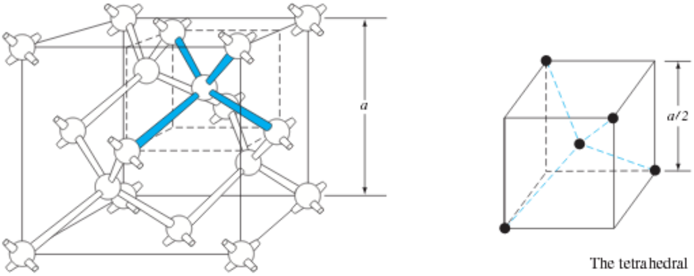
2.2.1.Solid State Equivalent of all the Vacuum Tube Devices.

Vacuum Tube Devices	Solid State Equivalent
Photo multiplier Tube	Photo diode/Op/.Amp.
Ruby and Gas LASER	PN Junction Diode LASER
Cathode Ray Oscilloscope	Flat Panel Display- LCD (Liquid Crystal Devices) and LED(Light-emitting Devices) version
T.V. Camera.	Charge-Couple Device Camera.
Thyratron	Thyristor or SCR(Silicon Contol Rectifier)
Vacuum Tube Triode and Pentode	Bipolar Junction Transistor(BJT), MOSFET(Metal-Oxide-

Semiconductor Field-Effect-Transistor),JFET(Junction-FET)

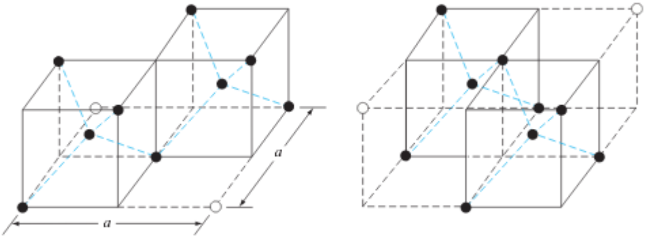
2.2.2. Crystalline Structure of Silicon

Silicon is a Group IV element with 4 electrons in its outermost orbit. Each Silicon atom is co-valently bond to 4 neighbouring Si atoms. In the process the octave is completed for all the Si stoms except those on the surface of the specimen. The surface has dangling bonds which need to be passivated. The crystalline structure of Silicon is called Diamond Lattice Structure.



The diamond structure.

The tetrahedral structure of closest neighbors in the diamond lattice.



(a)

(b)

Portions of the diamond lattice: (a) bottom half and (b) top half.

Figure 2.2.3. The Diamond Lattice structure of Silicon

Silicon has a lattice constant of ‘a’ = 5.43A° at 300K. It is two interpenetrating FCC(Face-Center-Cube) lattice with one sub-lattice displaced with respect to the other by quarter of the distance along the cross diagonal of the cube.

The whole diamond lattice cube is divided into 8 sub-cubes.Two sub-cubes in the upper half and two sub-cubes in the lower half contain the tetrahedral structure. The four sub-cubes are so arranged that the top view gives seamless continuity of co-valent sharing of the electrons with four neighbouring atoms as shown in Figure 2.2.4.

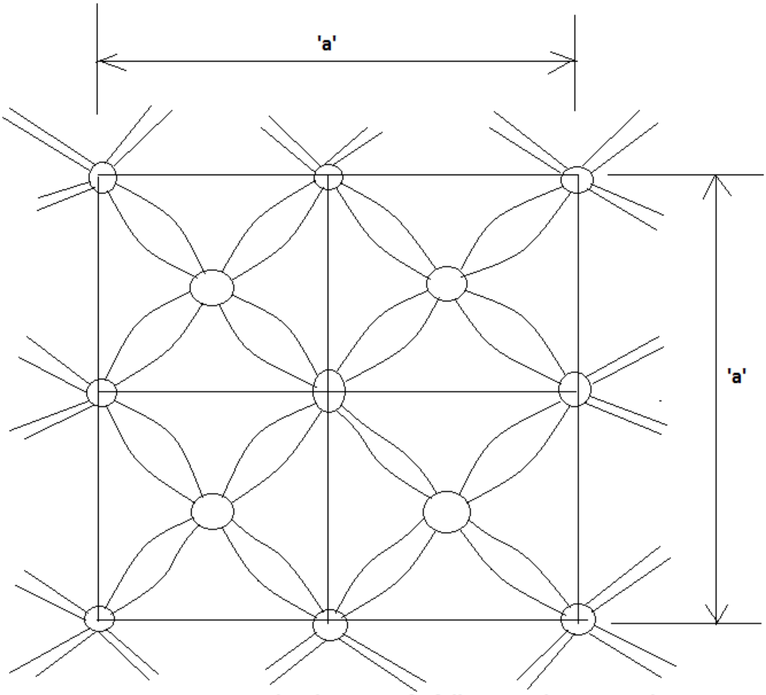


Figure 2.2.4.The Plan view of a full Diamond Lattice with cube edge of 'a' lattice parameter.

Tetrahedral structure shown in Figure 2.2.3. is repeadted throughout the lattice space as a consequence there are:

4 body atoms $= 4 \times 1 = 4$ atoms

6 face atoms $= 6 \times ½ = 3$ atoms.

8 corner atoms $= 8 \times 1/8 = 1$ atoms

Hence there are $4+3+1= 8$ atoms per unit cell.

Therefore atom number density $= 8 / a^3 = 8/(5.43\text{A}^\circ)^3 = 5 \times 10^{22}$atoms/cc.

Generally all crystalline solid material have a packing density of this order. This is the reason whuy in Copper each atom being univalent contributes 1 electron. Hence conducting electron concentration ‘n’ ~ 10^{22} electrons/cc .

Weight Density = atom density × weight per atom.

Weight per atom = Weight of 1 gram-mole/ Avogadro Number = A/A_{vog}

$$Weight\_Density = \rho = \frac{5\times10^{22}\,atoms/cc\times28.09\,gm/\mathrm{gm}-mole}{6.02\times10^{23}\,atoms/gm-mole} = 2.33\,gm/mole$$

Chapter2.2.3;2.2.4;2.2.5;Insulator,Semi-conductor,Metal;Intrinsic Semi-condctor;Extrinsic Semi-conductor.

Section 2.2.3. gives the distinction between Insulator, Semi-conductor and Metal.Section 2.2.4 gives thermal generation of Electron-hole pair in Intrinsic Semiconductor.Section 2.2.5 gives Extrinsic Semiconductor,Space Charge Neutrality and Compensation.

2.2.3. Distinction between Insulator,Semi-conductor and Metal based on the Band Theory of Solids.

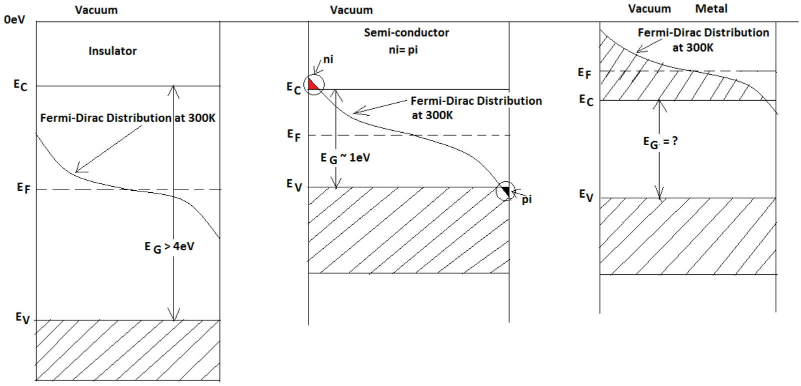


Figure 2.2.24. Comparative Band Diagram of Insulator, Semi-conductor and Metal.

In Figure 2.2.24. the comparative study of the Band-Diagrams of Insulator, Semi-conductors and Metal is given.Close examination of the Band-Diagrams reveal that there is really no difference between

Insulator and Semi-conductor except that Insulator has a band-Gap greater than 4eV and within the permissible range of temperatures, electrons can never jump across the band gap into the conduction band and Insulator will always be a bad conductor of electricity. In contrast in Semi-conductors, electrons from the valence band manage to jump across the band-gap right after crossing 77K temperature and at Room Temperature 300K there is appreciable amount of mobile electrons and holes as given in Table 2.2.4.1. Ge, Si and GaAs have intrinsic concentration of 10^{13}/cc, 10^{10}/ccand 10^{6}/cc respectively. That is precisely why Ge and Si behave as semi-conductor whereas GaAs behaves as semi-insulator. In contrast Metal always has conducting electrons in Conduction Band. Hence it is a good conductor at all temperatures. All metals have atom packing density of 10^{22}/cc and all metals are univalent hence every atom contributes 1 electron to the conduction band therefore metals in general have n(conducting electron concentration) = 10^{22}/cc. Metals generally have partially filled conduction band or overlapping valence and conduction band as shown in Figure 2.2.25.

There is one very distinctive difference between Semiconductor and Metal. As seen in Figure 2.2.24, Fermi-Level is in the middle of the Band Gap in semi-conductor whereas in Metal Fermi-Level lies in the conduction band and conduction band is filled up with conducting electrons till Fermi-Level.

As we will see later that as we dope a semiconductor, Fermi-level moves towards the Conduction Band in N-Type Semiconductor and towards the Valence Band in P-Type Semiconductor. In heavily doped semiconductors where doping level exceeds Effective Density of States at Ec in case of N-Type material and exceeds Effective Densitty of States at Ev in case of P-Type Semi-conductor then Fermi-level penetrates the Conduction Band N-Type Semiconductor and penetrates Valence Band in P-Type Semiconductor and they are called degenerate semi-conductor. These heavily doped semi-conductors behave like semi-metals.

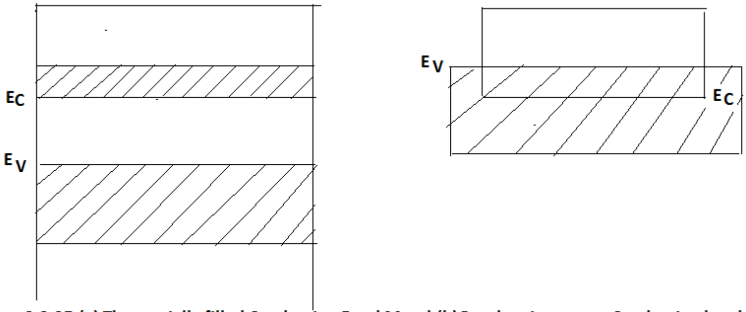


Figure 2.2.25.(a) The partially filled Conduction Band Metal;(b)Overlapping empty Conduction band and completely filled Valence band metal.

2.2.4.Thermal Generation of Elecron-Hole Pair(EHP) in Intrinsic Semiconductor.

Electronic Grade Semiconductor with no dopents are called Intrinsic Semiconductor. At temperatures below liquid Nitrogen temperature (77K), intrinsic semiconductor freeze out. There are no conducting electrons and no conducting holes. As temperature is raised above 77K, due to thermal vibration of the lattice structure a few covalent bonds break and n_i electrons leave their host atoms and by corollary p_i holes are created in the vacancies left by the semi-free electrons. Broken-loose electrons behave like negatively charged carriers with an effective mass near the mass of a free electron. The vacancies behave like positively charged carriers with effective mass heavier than that of negatively charged carriers. Here by definition the concentration of the two kind of mobile carriers are equal that is:

$$n_i = p_i = \sqrt{N_c N_v}\, Exp[-\frac{E_g}{2kT}] \_\_\_\_\_\_\_\_\_\_ (2.2.4.1)$$

Here N_C = effective density of states at E_C (the lower edge of Conduction Band) ~10^{19}/cc for all the three semiconductors namely Ge,Si and GaAs.

Here N_V = effective density of states at E_V (the upper edge of Valence Band) ~10^{19}/cc for all the three semiconductors namely Ge,Si and GaAs.

E_g = Band-Gap(eV), k = Boltzmann Constant, T = Absolute Temperature in Kelvin.

Table 2.2.4.1 gives the intrinsic concentration of electrons and holes in Ge,Si and GaAs at Room Temperature (300K).

Table 2.2.4.1. Intrinsic Concentration of electrons and holes in Ge,Si and GaAs at Temperature 300K.

Semiconductor	Band-Gap	Temperature	Effective Density of States	$n_i = p_i$
Ge	0.67eV	300K	~10^{19}/cc	2.4×10^{13}/cc
Si	1.12eV	300K	~10^{19}/cc	4×10^{9}/cc
GaAs	1.4eV	300K	N_C ~5×10^{17}/cc, N_V ~7×10^{18}/cc	3.3×10^{6}/cc

Figure 2.2.24 in the previous section illustrates how the EHP are thermally generated and why they are always equal.

Equation 2.2.3.1 clearly shows that intrinsic concentration is very temperature sensitive.

Table.2.2.4.2.Intrinsic Concentration of Si from 300K to 360K

Temperature	**300K**	**310K**	**320K**	**330K**	**340K**	**350K**	**360K**
Intrinsic Conc(per cc)	4×10^9	8×10^9	1.5×10^{10}	2.8×10^{10}	5×10^{10}	9×10^{10}	1.45×10^{11}

From Table 2.2.4.2 it is evident that for an increment of 60K above Room Temperature, intrinsic concentration has increased by two-orders of magnitude i.e. nearly 100 times. Because of this undesirable feature of intrinsic concentration, Si suffers from a very serious drawback namely: the reverse leakage current in a reverse biased diode doubles for every 10K rise in temperature.

In Figure 2.2.26, the exponential rise in intrinsic concentration with temperature is plotted.

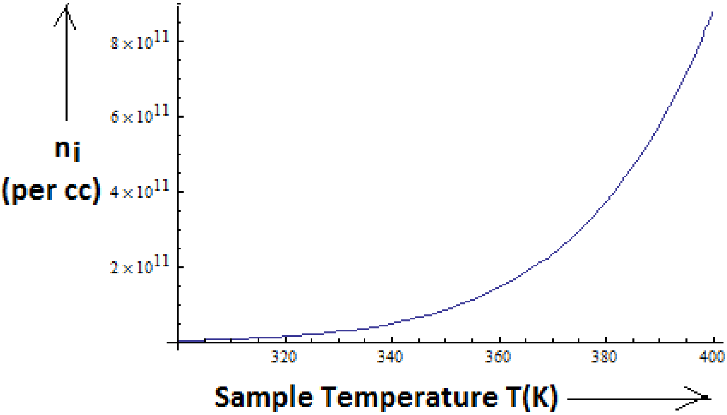


Figure 2.2.26.Exponential rise in Intrinsic Concentration of Si with the Sample Temperature.

2.2.5. Extrinsic Semi-conductor, space-charge neutrality and compensation .

Electronic Grade Semiconductor with no doping is called Intrinsic Semiconductor and that with doping is called Extrinsic Semiconductor.

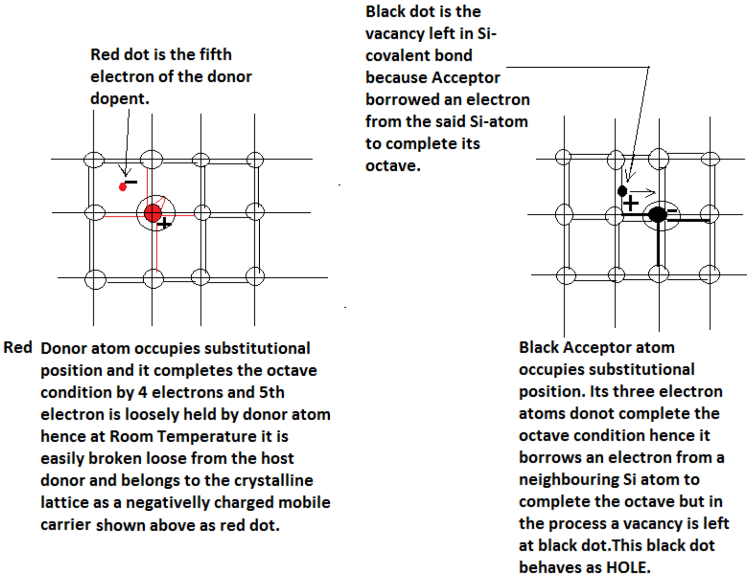
Semiconductor is Group IV element or compound semi-conductor is Group (III + V) Compound.

For Group IV element, Group III(namely B, Al) is acceptor or P-Type dopent and Group V(P, As) is donor or N-Type dopent.

For Group (III + V)Compound semiconductor, Group II is acceptor or P-Type dopent and Group VI is donor or N-Type dopent.

In Figure 2.2.27, the consequence of introducing a donor atom or the consequence of introducing an acceptor atom in a Silicon Crystalline Lattice is shown.

In case of Donor doped Silicon where doping density is N_D donor atoms/cc , the octave condition is completed by 4 valence electrons of donor atom hence 5th valence electron beomes the odd-man out and hence is very loosely held by the host donor atom. At Room Temperature, the 5th valence electron is broken loose from the donor and becomes a wanderer in Si-lattice. In effect Donor atom gets (+)ve ly ionized and donates its negatively charged 5th valence electron to the Si-lattice.Hence overall space-charge neutrality is maintained. Plus there are thermally generated EHP also.Overall thermal equilibrium is achieved and we have thermal equilibrium concentration value(n_n-bar) of conducting electron and thermal equilibrium value (p_n-bar) of holes. Here conducting electrons are many orders of magnitude larger than holes hence electrons are referred to as majority carriers and holes are referred to as minority carriers.


Black dot is the vacancy left in Si-covalent bond because Acceptor borrowed an electron from the said Si-atom to complete its octave.

Red dot is the fifth electron of the donor dopent.

Red Donor atom occupies substitutional position and it completes the octave condition by 4 electrons and 5th electron is loosely held by donor atom hence at Room Temperature it is easily broken loose from the host donor and belongs to the crystalline lattice as a negativelly charged mobile carrier shown above as red dot.

Black Acceptor atom occupies substitutional position. Its three electron atoms donot complete the octave condition hence it borrows an electron from a neighbouring Si atom to complete the octave but in the process a vacancy is left at black dot.This black dot behaves as HOLE.

Figure 2.2.27. The Physics of formation of N-Type and P-Type materials.

The thermal equilibrium value of the majority carrier n_n-bar and the thermal equilibrium value of the minority carrier p_n-bar are calculated from the following set of equations:

$$By\_Law\_of\_Mass\_Action\_\overline{n_n}\times\overline{p_n}=n_i^2\_and\_By\_Space\_Charge\_Neutrality\_\overline{n_n}^{(-)}=N_D^{(+)}+\overline{p_n}^{(+)}\_\_\_(2.2.5.1)$$

By re-arranging the above two equations we get the following equation:

$$\bar{n}_n \times (\bar{n}_n - N_D) = n_i^2;$$

Opening_the_brackets_we_get_a_quadratic_equation:

$$\bar{n}_n^2 - \bar{n}_n N_D - n_i^2 = 0;$$

The_solution_of_the_quadratic_equation_is:

$$\bar{n}_n = \frac{N_D \pm \sqrt{N_D^2 + 4n_i^2}}{2};$$

Rearranging_the_above_equation_we_get:

$$\bar{n}_n = \frac{N_D \pm N_D \sqrt{1 + \frac{4n_i^2}{N_D^2}}}{2};$$

If $_(\frac{2n_i}{N_D})^2 << 1$_then_Binomial_Expansion_is_applicable:

$$\bar{n}_n = \frac{N_D \pm N_D (1 + \frac{1}{2} \times \frac{4n_i^2}{N_D^2})}{2} = \frac{1}{2} N_D \pm \frac{1}{2} N_D \pm \frac{n_i^2}{N_D};$$

Hence_thermal_equilibrium_value_of_majority_carriers_has_two_values:

$$\bar{n}_n = N_D + \frac{n_i^2}{N_D} \text{ and\_} \bar{n}_n = -\frac{n_i^2}{N_D};$$

The_negative_value_is_untenable_so_we_retain_the_positive_Value:

$$\bar{n}_n = N_D + \frac{n_i^2}{N_D} \text{\_and\_} \bar{p}_n = \frac{n_i^2}{\bar{n}_n};$$

But_for_all_practical_purposes:

$$\bar{n}_n = N_D; \text{Therefore\_} \bar{p}_n = \frac{n_i^2}{N_D}$$

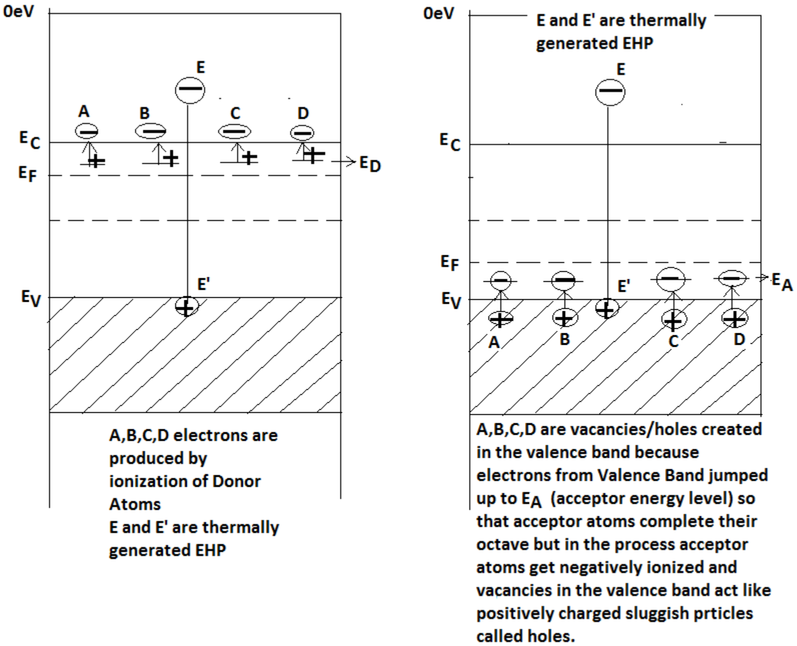
In second last equation we see that majority carriers consist two terms: first term is electrons contributed by ionization of the ND dopent atoms/cc and the second term is the electrons contributed through thermal generation of EHP. But since electrons contributed by EHP thermal generation is negligible compared to ND dopent atoms/cc as seen below so without any loss of accuracy we can use the last equation above.

In case of Silicon, doping is in the range from 10^{12}Dopent Atoms/cc to 10^{18}Atoms/cc.

Hence the thermal equilibrium value of the majority carrier in N-Type Si will range from n_n-bar = 10^{12}electrons/cc to 10^{18} electrons/cc.

The thermal equilibrium value of the minority carier in N-Type will range from p_n-bar = 10^8holes/cc to 10^2 holes/cc. assuming $n_i = 10^{10}$/cc.

2.2.5.1. The Energy-Band Diagram of N-Type and P-Type Semiconductor.



A,B,C,D electrons are produced by ionization of Donor Atoms
E and E' are thermally generated EHP

E and E' are thermally generated EHP

A,B,C,D are vacancies/holes created in the valence band because electrons from Valence Band jumped up to E_A (acceptor energy level) so that acceptor atoms complete their octave but in the process acceptor atoms get negatively ionized and vacancies in the valence band act like positively charged sluggish prticles called holes.

Figure 2.2.28. Energy Band Diagram of N-Type and P-Type Silicon Sample.

Left hand side diagram in Figure 2.2.28 is that of N-Type Material. The dopent N_D donor atoms/cc introduce filled energy level E_D very close to the lower edge (E_C) of the Conduction band.At room temperature the thermal energy is of the order of 25meV whereas 10meV is required to jump to the conduction band hence at Room Temperature all the donor atoms easily get positively ionized contributing their 5th valence electrons to the conduction band. There are few and far between thermal generation of EHP also as indicated by EE'.

Right hand side diagram in Figure 2.2.28 is that of P-Type Material. The dopent N_A acceptor atoms/cc introduce empty energy level E_A very close to the upper edge (E_V) of the Valence band.At room temperature the thermal energy is of the order of 25meV whereas 10meV is required to jump to the empty energy level E_A hence at Room Temperature all the acceptor atoms easily get negatively ionized by accepting the valence electron from the neighbouring Si-Atom thereby completing its octave and contributing equal number of vacancies in the Valence Band. These vacanicies wander about the Si-lattice as conducting positive carriers which we call holes. There are few and far between thermal generation of EHP also as indicated by EE'.

Comparing these two diagrams it is clear that holes will wander about the Si-lattice in much more sluggish manner as compared to conducting electrons. This has implications for its mobility.

2.2.5.2. Compensation in Semiconductors.

If both donor N_d dopent/cc and acceptor N_a dopent/cc are introduced simultaneously then the two compensate. The fifth electron from the donor atom goes to fulfil the octave of the acceptor atom. The Acceptor need not make its pick of the 4th electron from the valence band. Thus equal number of donor and acceptor will compensate each other.Whatever is leftover decides the Type of the semiconductor.

If donor N_d dopent/cc $<$ acceptor N_a dopent/cc then we have P-Type Si with net doping density $= N_a - N_d = N_A$.

If donor N_d dopent/cc $=$ acceptor N_a dopent/cc then we have Intrinsic Si with net doping density $= N_a - N_d = 0$.

If donor N_d dopent/cc $>$ acceptor N_a dopent/cc then we have N-Type Si with net doping density $= N_d - N_a = N_D$.

SSPD_Chapter_2.2.6. Theoretical Formulation of thermal equilibrium values of majority and minority carriers in Semi-conductors.

SSPD_Chapter 2.2.6 gives the theoretical formula for determining the majority and minority carrier density.

2.2.6. Theoretical Formulation of thermal equilibrium values of majority and minority carriers in Semi-conductors.

The broad view of Silicon(intrinsic or doped) is that there is thermal generation of EHP(electron-hole pair), EHP are recombining and there is the ionization of the dopent.At any given temperature we have a thermal equilibrium established. At this thermal equilibrium, the concentrations achieved are time invariant and are referred to as thermal equilibrium concentrations.

In intrinsic Silicon: electron thermal equilibrium concentration = hole thermal equilibrium concentration = intrinsic concentration.

$$\bar{n} = \bar{p} = n_i$$

The bar indicates thermal equilibrium value.

In N-Type Silicon we have:

$$Majority\_carrier = \overline{n}_n, Minority\_carrier = \overline{p}_n;$$

In P-Type Silicon we have:

$$Majority\_carrier = \overline{n}_p, Minority\_carrier = \overline{p}_p;$$

By Law of Mass Action in each case:

$$Majority\_carrier \times Minority\_carrier = n_i^2;$$

2.2.6.1. Majority and Minority Carrrier thermal equilibrium concentrations as a function of Temperature in Kelvin.

In intrinsic Silicon, thermal equilibrium concentrations of electrons and holes are equal and are called intrinsic concentration. In Figure 2.2.26. the Intrinsic Concentration of Silicon as a function of Temperature(K) is given. It should be noted that conducting electron and holes in intrinsic Si is only due to thermal generation of EHP. Hence below 155K(liquid Nitrogen temperature), thermal generation stops and intrinsic Si becomes a perfect insulator.

In Figure 2.2.29. Fermi-Dirac distribution is given at T=155K and at T=300K. At T= 300K, Fermi-Dirac distribution is more skewed hence the tail end of the distribution has a greater distribution as compared to that at 155K.

At 300K, $n_i = 4\times10^9$ and $P(E_C) = 4\times10^{-10}$;

At 155K, $n_i = 6.55$ and $P(E_C) = 6.55\times10^{-19}$;

For all practical purposes, thermal generation of EHP has altogether stopped and the sample freezes out. It becomes an insulator.

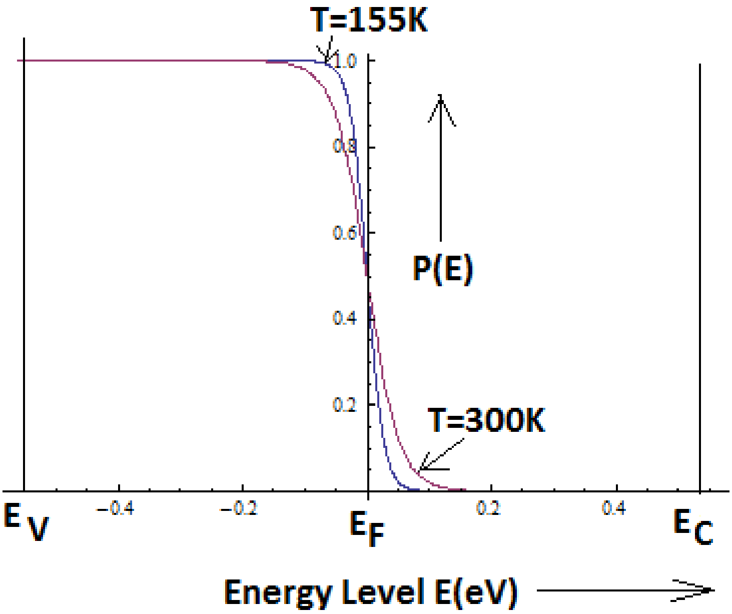


Figure 2.2.29. Fermi-Dirac Distribution at T = 155K and at T=300K.

Blue curve is 155K and pink curve is 300K

In Extrinsic Silicon, below 155K, there is no thermal generation of EHP as well as there is no ionization of the dopent. Hence there are no charge carriers whatsoever in Si below 155K.

From 155K to 200K, there is thermal generation of EHP but there is no ionization because the thermal lattice vibration does not provide enough energy for ionization. Above 200K the thermal energy is kT = 17meV whereas ionization energy is only 10meV hence practically all dopent atoms get ionized and majority and minority carriers remain constant inspite of the fact that EHP thermal generation is exponentially increasing. This increase is felt only above 473K when intrinsic carrier concentration starts dominating the dopent concentration.

In extrinsic Si, dopent concentration is in the range 10^{12}/cc to 10^{15}/cc whereas the intrinsic concentration at 473K and above is above 10^{13}/cc.

Well above 473K depending on the dopent concentration, the sample becomes intrinsic type.

The thermal equilibrium values of majority and minority carriers in N-Type Silicon with a donor doping density of 10^{13}/cc as a function of temperature is given in Figure 2.2.30.

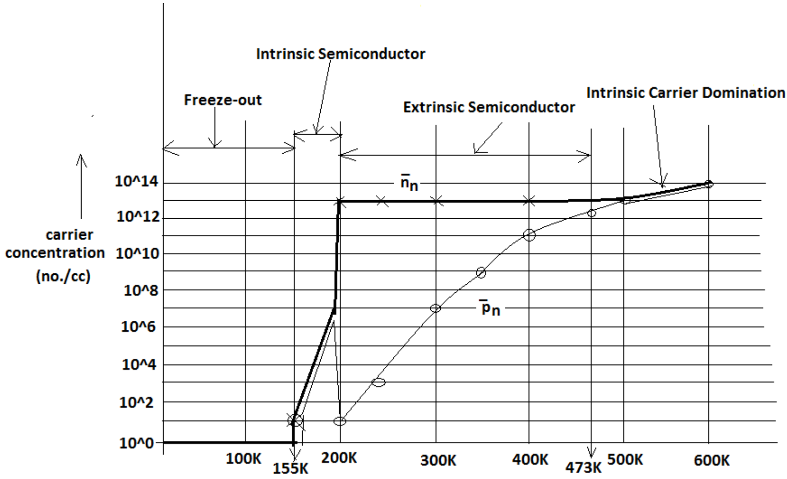


Figure 2.2.30. Majority and minority carrier thermal equilibrium concentration variation with temperature in N-Type Silicon with Donor Dopent density=10^13/cc.

There are four distinct regions of Doped Silicon:

Freeze out region (0K to 155K) - neither there is EHP thermal generation nor there is the ionization of the donor atoms.

Intrinsic Semiconductor (155K to 200K) - thermal generation of EHP starts but there is no ionization of Donor Atoms.

Extrinsic Semiconductor (200K to 473K) – Ionization of Donor atoms is 100% hence majority carriers thermal equilibrium density is decided by Donor Density hence it is constant at 10^{13}/cc. Thermal generation of EHP is taking place but it is many orders of magnitude less than the majority carrier hence from 200K to 473K resistivity of the sample is relatively constant.

Intrinsic Carrier Concentration dominates over the doping density(473K and beyond) – at 473K , thermal generation of EHP becomes comparable to doping density. At 600K it becomes one order of magnitude greater than the doping density. So in this region it again behaves as Intrinsic Semiconductor.

2.2.6.2.Theoretical Formulation of Majority Carriers and Minority Carriers.

Because of Pauli-Exclusion Principle, we have a term called Density of States N(E). This essentially means the number of electrons which can be accommodated per unit volume per unit eV in permissible energy states.

In the Appendix of this Chapter I will show that:

$$N(E)=\frac{dn}{dE}=\frac{8\pi\sqrt{2m_e}}{(h)^3}m_e\sqrt{E}=\frac{3.392339521\times10^{-14}}{(6.62606876\times10^{-34})^3}\times9.10938188\times10^{-31}\sqrt{E}=1.0621677123446\times10^{56}\sqrt{E}$$

(2.2.6.2.1)

N(E)dE=number of energy states per m3 between energy level E and E+dE.

Therefore actual number of electrons present in conduction band between energy level E and E+dE are = P(E)N(E)dE

Where P(E) = 1/[1+Exp[(E-EF)/kT];

But at T $=$ 0K, P(E) $=$ 1 upto E $=$ E_F hence

actual number of electrons present in conduction band between energy level E and E $+$ dE are $=$ (1)N(E)dE

In Copper we know that electrons are filled up to the Fermi-Level and that the Fermi-level is at 7.05eV as shown in Figure 2.2.31.

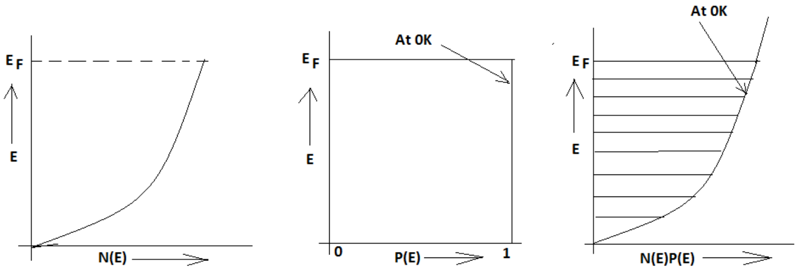
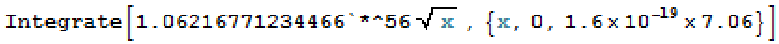


Figure 2.2.31.Density of states, Fermi-Dirac Distribution at 0K and the occupancy of permissible energy states in the conduction band of Copper.

Hence number of conducting electrons in the conduction band from the bottom of the conduction band to Fermi-level $= \int (1)K\sqrt{E}dE$, $\{E, 0, 7.05 \times 1.6 \times 10^{-19}\}$

Here K is a constant $=$ $1.06216771234466 \times 10^{56}$.

Using Mathematica:



```
Integrate[1.06216771234466`*^56 √x , {x, 0, 1.6×10^-19 ×7.06}]
```

(2.2.6.2.2)

Conduction Band conducting electron density = n= $8.5\times10^{28}/m^3 = 8.5\times10^{22}/cc$.

This also implies that Copper atom packing density in Copper metal is $8.5\times10^{28}/m^3 = 8.5\times10^{22}/cc$.

You can see normally the packing density is of this order i.e.$10^{22}/cc$.

Exactly the same procedure is applied to determine the majority carrier thermal equilibrium density.

In N-Type Silicon:

Number of electrons actually present in the conduction band at 300K= ∫P(E)N(E)dE .

This is a definite integral between the lower limit E_C and the upper limit infinity as shown below:

$$n=\int_{E_c}^{\infty} N(E)P(E)dE=\int_{E_c}^{\infty}\frac{K\sqrt{E-E_c}}{1+Exp[\frac{E-E_F}{kT}]}dE;$$

(2.2.6.2.3)

It will be shown in the Appendix of this Section that this integral comes out to be:

Electron density= n_{n---} = N_CExp[-(E_C-E_F)/(kT)]__________(2.2.6.2.4)

Where $N_C = 2[(2\pi m_e^*kT)/h^2]^{3/2} = 2.82 \times 10^{25}/m^3 = 2.82 \times 10^{19}$/cc.

Here m_e^* = effective mass of the electron at the bottom of the conduction band.

For non-degenerate semiconductors where E_F is several kT below E_C , in such cases

Fermi-Dirac Statistics($P(E) = 1/[1+\text{Exp}[(E-E_F)/kT]]$) approximates Boltzmann-Maxwell Statistics namely $P(E) = \text{Exp}[-(E-E_F)/kT]$.

At $E = E_C$ this reduces to $P(E_C) = \text{Exp}[-(E_C - E_F)/kT]$.

Hence in Equation 2.2.6.2.4, if $\text{Exp}[-(E_C - E_F)/kT]$ is the probability of occupancy at the lower edge of the conduction band and n_{n}--- is the conduction electron density in conduction band then N_C is the effective density of states at E_C.

We follow the same procedure for holes concentration calculation:

Number of holes actually present in the valence band at 300K $= \int(1-P(E))N(E)dE$

This is a definite integral between the lower limit E_V and the upper limit (- infinity) as shown below:

$$p = \int_{E_v}^{-\infty} N(E)(1-P(E))dE = \int_{E_v}^{-\infty} \frac{K\sqrt{E_v - E} \times Exp[\frac{E-E_F}{kT}]}{1+Exp[\frac{E-E_F}{kT}]} dE;$$

(2.2.6.2.5)

As will be shown in the Appendix this reduces to:

Hole density = p_{n---} = $N_V Exp[-(E_F-E_V)/(kT)]$ __________(2.2.6.2.6)

Where $N_V = 2[(2\pi m_h^* kT)/h^2]^{3/2} = 1.83 \times 10^{25}/m^3 = 1.83 \times 10^{19}/cc$.

Since $Exp[-(E_F-E_V)/(kT)]$ = probability of non-occupancy at E_V = probability of occurrence of hole at the upper edge of the valenvce band hence N_V = is the effective density of states at the upper edge of the valence band.

$N_V = N_C$ this identity does not hold since the effective masses are not the same as seen in Table 2.2.6.1.

By Law of Mass Action: $n_{n---} \times p_{n---} = n_i^2 = N_C N_V Exp[-E_G/(kT)]$ where $E_G = E_C - E_V$;

Hence intrinsic carrier concentration =

$n_i = \sqrt{(N_C N_V)} \times Exp[- E_G/$

(2kT)]________________(2.2.6.2.7)

In Table 2.2.5.1. a comparative study of the band-gap, effective mass of electron and hole, effective density of states and intrinsic carrier concentration is made for Ge,Si and GaAs.By the inspection of Table 2.2.5.1, it is evident that density of states are in the range of 10^{19} per cc for all three semi-conductors..

Table 2.2.6.1. Band-gap, effective mass , effective densiy of states and intrinsic carrier concentration at 300K.

All parameters at 300K	Ge	Si	GaAs
E_G(eV)	0.66	1.12	1.424
(effecteive mass)m_e^* ($\times m_e$)	0.56	1.08	0.067
(effecteive mass)m_h^* ($\times m_e$)	0.29	0.81	0.47
N_C(per cc)	1.05×10^{19}	2.82×10^{19}	4.37×10^{17}
N_V(per cc)	3.92×10^{18}	1.83×10^{19}	8.68×10^{18}
n_i(per cc)	1.83×10^{13}	8.81×10^{9}	2.03×10^{6}

SSPD_Chapter_2.2.6. Drift Velocity in Semiconductor/metal and its conductivity.
SSPD Chapter 2.2.7.gives the proof of Ohm's Law and gives the temperature dependence of resistance in Semiconductor and Metal.

2.2.7. Drift Velocity in Semiconductor/ metal and its conductivity.

In conductors it is the drift of the electrons which gives rise an electric current. It is the drift of the electrons which gives rise to Ohm’s Law (V/I = R). The flow of electrons in an electric circuit is shown in Figure 2.2.32.

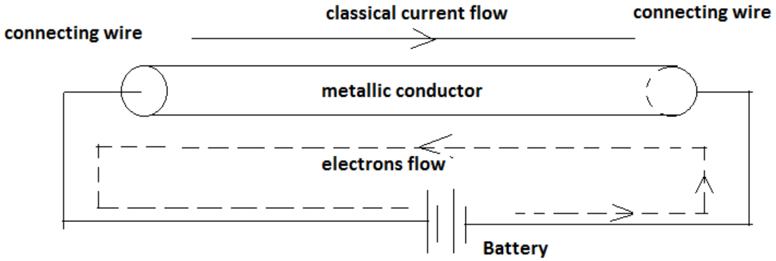


Figure 2.2.32. Flow of electrons in an electric circuit.

Notice here there is only electron flow from the negative terminal to the positive terminal of the battery and right from the middle of 18th century we have been mistakenly calling it a current flowing from positive to the negative terminal as if it is a flow of some positive charges. In an electric circuit there is the flow of electrons only.

The electric conductor obeys Ohm’s Law:

$$R = \frac{V}{I} = \rho \times \frac{l}{A} = \frac{l}{\sigma \times A} ohms\_\rho = \mathrm{Resistivity}(\Omega - cm)\_\sigma = Conductivity(Siemens/cm\_or\_mho/cm)$$

(2.2.7.1)

A = cross-sectional area of the conductor, l = length of the conductor.

In contrast semiconductor allows the current to be carried by electrons as well as by holes. Hence electric current is partly carried by electrons and partly carried by holes as shown in Figure 2.2.33.

As shown in Figure 2.2.33. since the Si Rod is intrinsic in nature, 50% of current is carried by electrons and 50% is carried by holes within the Si-Rod. In external circuit, 100% of the current is carried by electrons sourced from the negative terminal of the battery and drained out by the positive terminal of the battery. At the negative terminal all the holes flowing in the Si-Rod recombine with a part of the electrons brought by the external circuit. In this particular example 50 holes/sec recombine with 50 electrons/sec. At the positive terminal equal number of electrons and holes are generated and maintain the continuity of the current across the positive interface.In this example 50 EHP/sec are generated and maintain the continuity of current.

If the Si-Rod is extrinsic then major portion of the current is carried by the majority carriers and minor portion of the current is carried by the minority carriers.

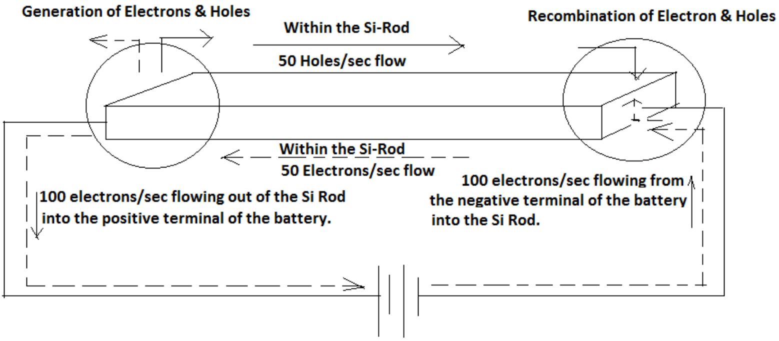
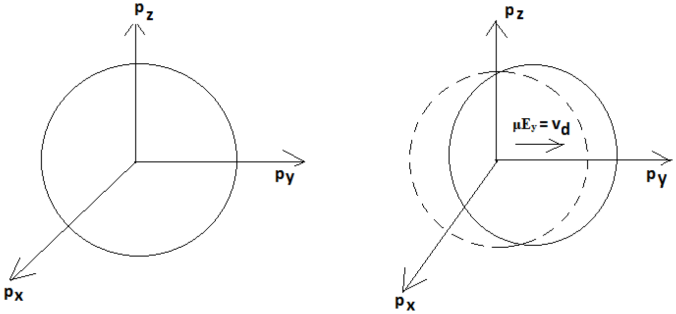

Generation of Electrons & Holes
Within the Si-Rod
50 Holes/sec flow
Recombination of Electron & Holes
Within the Si-Rod
50 Electrons/sec flow
100 electrons/sec flowing out of the Si Rod into the positive terminal of the battery.
100 electrons/sec flowing from the negative terminal of the battery into the Si Rod.

Figure 2.2.33. Current Flow through a Silicon Rectangular Rod of intinsic nature.In the rod half the current is carried by the holes and half the current is carried by electrons.

In Chapter 1, Part 12, the quantum mechanical interpretation of Resistance of a conductor is given.



Without applied Electric Field:
In momentum space, we get a sphere centered at the origin.The thermally agitated conducting electrons which are moving randomly have no net motion.

With applied Electric Field in Y-direction:
There is a net electric drift velocity v_d in Y-direction. Hence momentum vector sphere representing all the conducting electrons is displaced with respect to the origin in Y-Direction.
As long as v_d << Vth (10^7cm/s=10^5m/s) we have quasi equilibrium and ohm's law is satisfied. Extra energy given during the drift is given up to the crystal as vibrational energy. Hence drift velocity is directly proportional to the Electric Field.

Figure 2.2.34. Drift motion is a perturbation in the thermal motion with quasi-equilibrium condition.

As seen in Figure 2.2.34. the drift velocity (v_d) is a perturbation in the random thermal motion of the conducting electrons under an applied electric field E_Y. The otherwise centrally placed momentum sphere gets displaced in Y-direction by v_d under the application of an Y-directed electric field.

The thermal velocity of electron or hole in Si-Rod is given by the following equation:

$$\frac{3}{2}kT(thermal\_energy\_by\_equipartiition\_law)=\frac{1}{2}m_e v_{ther}^2;$$

$$Therefore\_v_{ther}=\sqrt{\frac{3kT}{m_e}}=1.168\times10^5 m/s(at\_300K);$$

(2.2.7.2)

In Chapter 12, Part 12, we saw that electron or hole gets accelerated in spurts. At short intervals of time it loses all energy to the crystal and its motion is reset. That is it restarts from zero velocity. This will happen as long as drift velocity < < thermal velocity.Under this condition:

Drift Velocity is directly proportional to the applied Electric Field and v_d□$_Y = \mu E_Y$. The constant of proportionality is called mobility μ.

We will show that under this direct proportionality condition Ohm’s Law is satisfied.

2.2.7.1. Mobility of electron in Semiconductor is intimately connected with the structure of the Semi-conductor.

In Chapter 12, Part 12, it hasa been shown that:

$$\mu = \frac{q\tau}{2m_e^*}\_where\_\tau = mean\_free\_time;$$

(2.2.7.3)

Here mean free time is a critical parameter which depends on three factors related to the structure of the Semi-conductor.:

1. Ambient Temperature which causes the lattice vibration;
2. The impurity in the semi-conductor caused during ther growth of the single crystal;
3. The defects of the semi-conductor also caused during the growth of the single crystal.

If there were no disturbances(either due to thermal vibrations or due to defects in the crystalline structure or due to interstitial or substitutional impurities) in the orderly arrangement of the lattice centers then a mobile carrier under the application of an electric field would move in a straight line and would continue to move in straight line indefinitely even after the Electric Field is removed. This would be a living embodiment of the First Law of Newton:

A body at rest would continue to be at rest and a body in motion would continue to move in a straight line with a constant velocity unless or until made to act otherwise by the application of a force.

Such an ideal crystal would behave like a Super-conductor and its mean free time would be very large and its mobility would also be very large. .

In real life we always face disorderliness due to thermal vibrations, due to lattice defect and due to impurities. This leads to scattering and this lead to finite mobility and finite resistance. Hence we have two kinds of scattering and two kinds of mobilities:

$$\mu_{lattice}\_due\_to\_lattice\_scattering\_and$$

$$\mu_{imp}\_due\_to\_impurity\_scattering\_\text{and\_defect\_scattering\_and}$$

$$\frac{1}{\mu_{effective}}=\frac{1}{\mu_{lattice}}+\frac{1}{\mu_{imp}};$$

$$\mu_{lattice}\_\propto\_(\frac{1}{T^{3/2}});$$

$$\mu_{imp}\_\propto\_(\frac{T^{3/2}}{N_{imp}});$$

(2.2.7.4)

Using this theoretical formulation and experimentally determined mobilities, we can derive mean free time and gain useful insight in the structure of metals and semiconductors.

From Chapter 12, Part 11, we have taken the following Tables:

Table 2.2.7.1. Mean Free Time(τ) and Mean Free Path(L) in Metals.

Metal	μ_n(cm^2/(V-s))	E_F(eV)	V_e(10^5/s)	τ(femtosecond)	L(Angstrom) =
Li	16.2	4.7	9.96	9	90
Na	54.17	3.1	8.08	31	250
K		2.1	6.65	44	293
Cu	47	7.0	12.15	27	328
Ag	72	5.5	10.77	41	441.6

Table2.2.7.2. Mean Free Time and Mean Free Path in Semiconductors.

Material	μ_n(cm^2/(V-s))	τ(femosecond)	V_e(10^5/s)	L(A°)
Ge	3900	2217	0.95	2106
Si	1350	767.6	0.95	729
GaAs	8600	4890	0.95	4645.5

Electron is a matter wave hence it has a de-Broglie Wave-length associated with it which will decide how a conducting electron iinteracts with its surrounding lattice structure.The method of arriving at de-Broglie is described in Chaptetr 1,Part 11.

Table 2.2.7.3.de-Broglie Wavelength of electron in Electron Microscope, Metal and Semi-conductor and its implications for mobilities.

	V_{acc}(Volts)	**De-Broglie λ(m)**	**V_e(m/s)**	**Remark**
(i)Electron-μScope	10KV	10^{-11} = (1/50)× lattice Constant	5×10^{9}	Wave-length < < a
(ii)Conducting electron in Metal	4V	5×10^{-10} = lattice Constant	10^{6}	Wave-length ~ a
(iii)Conducting electron in Si	0.025V	6×10^{-9} = 12× lattice Constant	10^{5}	Wave-length > > a

Case i.In an Electron-Microscope, a regular crystal array does not scatter the electron beam. Hence an electron beam can penetrate a thin specimen and produce the image of the broad features without being influenced by the atomic details.

Case ii.In a metal, lattice centers strongly scatter the conducting electron hence their mobility is very low.

Case iii.In a semi-conductor or doped semi-conductor, conducting electrons have a wave-length of 60A°. Hence only gross form of imperfections, extending over several lattice constants such as lattice vibration or dislocations, cause scattering of the conducting electron. Hence mobility is much higher in Si,Ge or GaAs as compared to Metal.

2.2.7.2. Theoretical Formulation of Ohm’s Law.

Calculation of the drift current (J_{drift}) under the application of an Electric Field.

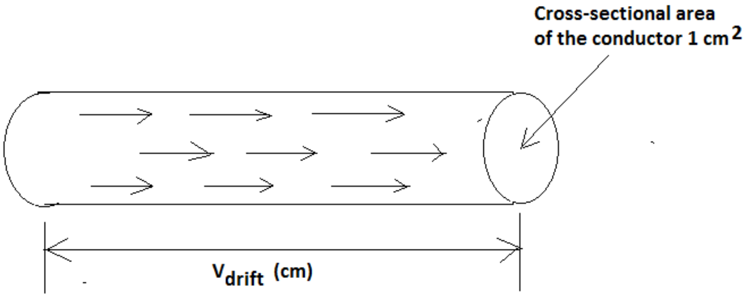


Figure 2.2.35.The model for calculating the drift current.

Refer to Figure 2.2.35.

J_{drift}□e(drift current due to holes) = number of coulombs of electron charge passing through the unit cross-sectional area per unit time

$= q \times n(\text{electrons/cc}) \times v_{\text{drift}}(\text{cm}) \times 1(\text{cm}_2)$

Therefore $J_{\text{drift}\square\text{e}} = qnv_{\text{drift}\square\text{e}}$(2.2.7.5)

$J_{\text{drift}\square\text{hole}} = qnv_{\text{drift}\square\text{hole}}$(2.2.7.6)

Total Drift Current $J_{\text{Drift}} = qnv_{\text{drift}\square\text{e}} + qnv_{\text{drift}\square\text{hole}}$(2.2.7.7)

$J_{\text{Drift}} = qn\mu_n E + qn\mu_p E = (qn\mu_n + qn\mu_p)E = \sigma E$ (2.2.7.7)

Where conductivity of the Si-Rod σ(Siemens/cm) = $(qn\mu_n + qn\mu_p)$(2.2.7.8)

But $\sigma = 1/\rho$ and $E = V/l$ where V = applied Voltage across the sample and l = length of the sample.

Therefore Eq.(2.2.7.7) is rewritten as:

I_{drift}/A(cross-sectional area of the Si-Rod) $= (1/\rho) \times (V/l)$(2.2.7.9)

Hence Eq.(2.2.7.9) is rearranged as:

$$I_{drift} = \frac{V}{\rho \times \frac{l}{A}} = \frac{V}{R}\_\,and\,\_R = \rho \times \frac{l}{A};$$

(2.2.7.10)

Eq.(2.2.7.10) is in fact Ohm’s Law.

2.2.7.3. Scatter Limited Velocity.

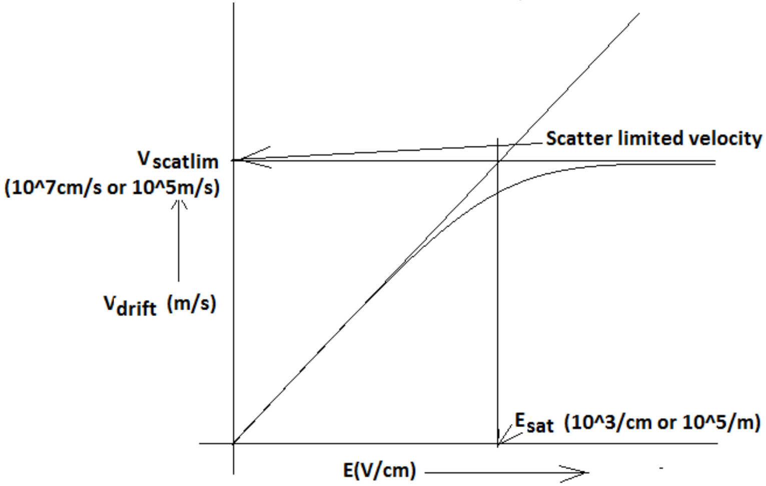


Figure 2.2.36. Plot of Drift Velocity with respect to Electric Field.

At low fields, after every scattering electron manages to transfer its excess energy to the lattice and electron gets reset. But as drift velocity approaches Thermal Velocity $=\surd(3kT/m_e)=10^7$cm/s the quasi equilibrium position becomes unachievable and electron’s drift velocity saturates at scatter-limited velocity which is same as the thermal velocity 10^7cm/s. This occurs at E $= 10^3$/cm.

From this point onward Ohm’s law is untenable.

2.2.7.4. Temprature dependence of conductivity of metal and semi-conductor.

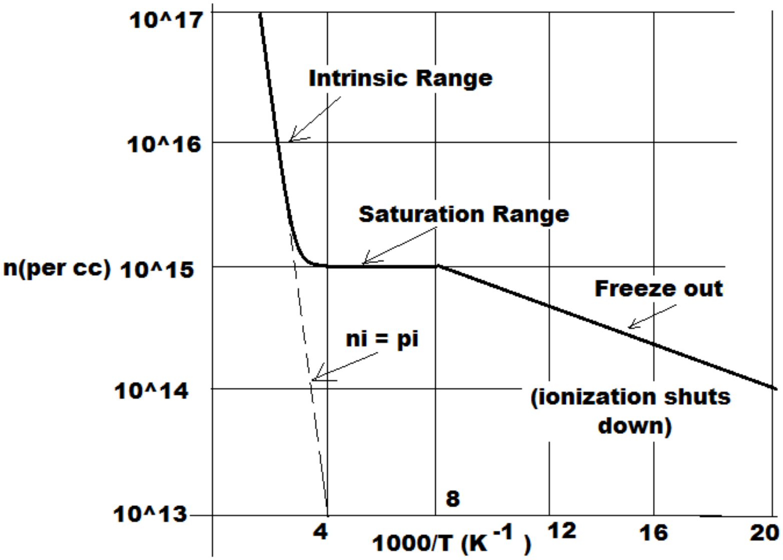


Figure 2.2.37.Electron Density as a function of reciprocal of Temperature for a Si-Sample wth doping density of N_D =10^15/cc

Using Eq.(2.2.7.4) the dependence of mobility is as shown in Figure 2.2.38.

In Figure 2.2.39. the temperature dependence of resistivity of Extrinsic Semiconductor and Metal are shown. By inspection of the Figure we find that Metal has a positive temperature coefficient of resistance whereas Semiconductor has negative temperature coefficient of resistance in intrinsic dominated region. In freeze out region its resistivity approaches that of an insulator.

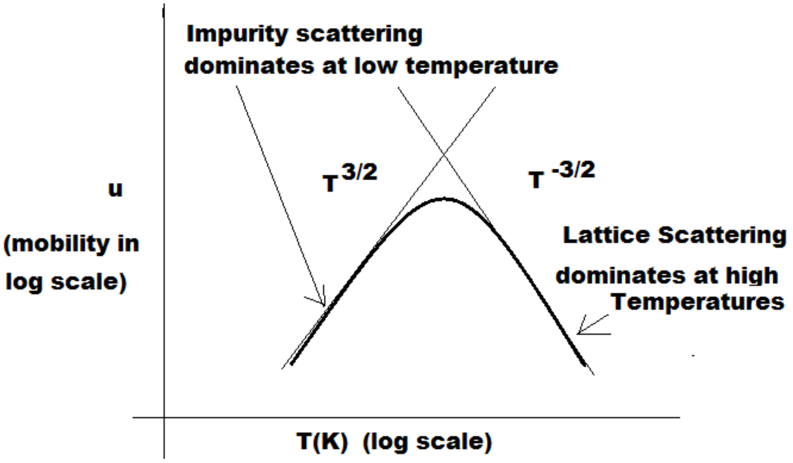


Figure 2.2.38.Temperature dependence of mobility in Si.

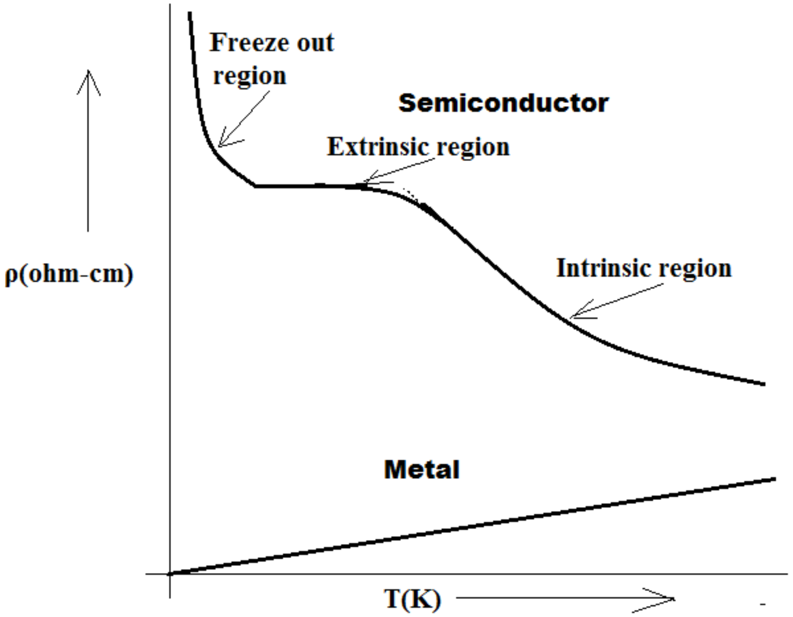


Figure 2.2.39. Temperature dependence of resistivity of Semiconductor and metal.

SSPD_Chapter_2.2.9-2.2.10-2.2.11:Diffusion,Life-Time and Continuity Equation.

Chapter 2.2.9,2.2.10 and 2.2.11 gives the diffusion process, recombination process and carrier concentration profile under steady state condition for optically perturbed Si-Bar.

2.2.9.Diffusion Process and Diffusion Current.

In kinetic theory of gases we have studied the concept of diffusion. Gas tends to move from higher concentration to lower concentration until gas is uniformly distributed through the room.

Just as potential gradient causes drift current in a similar manner concentration gradient causes a diffusion current. Linear concentration gradient causes a constant diffusion current as shown in Figure 2.2.40.

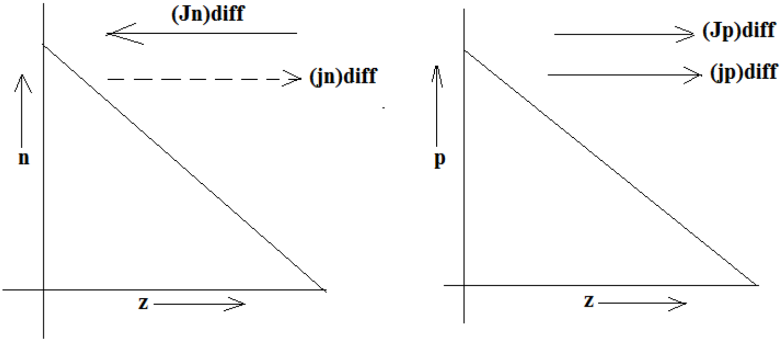


Figure 2.2.40. Illustration of electron diffusion current and hole diffusion current.

Conducting electrons and holes always move from high concentration to low concentration hence the rate of carrier movements per unit area or carrier flux density is proportional to the negative of the concentration gradient as given in the equation below:

$$(j_n)_{diff} = electron\_flux\_density = number\_of\_electrons\_passing\_through\_1(\mathrm{cm})^2\, per\_sec = -D_n \frac{\partial n}{\partial z};$$

$$(j_p)_{diff} = hole\_flux\_density = number\_of\_holes\_passing\_through\_1(\mathrm{cm})^2\, per\_sec = -D_p \frac{\partial p}{\partial z};$$

(2.2.9.1)

Here D_n and D_p are diffusion coefficient of electron and holes respectively.

The unit of Diffusion Coefficient = cm^2/s and Einstein Relation gives the inter-relation between Diffusion Coefficient and mobility namely:

$$\frac{D_n}{\mu_n} = \frac{D_p}{\mu_p} = \frac{kT}{q} = 26mV\_at\_300K;$$

(2.2.9.2)

Table 2.2.9.1. tabulates the values of diffusion coefficient and mobility foe conducting electrons and holes in intrinsic Si at 300K.

Table 2.2.9.1. Comparative values of diffusion coefficients and mobilities of conducting electrons and holes in intrinsic Si at 300K.

Material	μn(cm2/V-s)	μp(cm2/V-s)	Dn(cm2/s)	Dp(cm2/s)
Ge	3900	1900	100	50
Si	1350	480	35	12.5
GaAs	8500	400	220	10

When we assign the electronic charge to the flux density we obtain diffusion current density due to conducting electron or due to holes.

$$(J_n)_{diff} = +qD_n \frac{\partial n}{\partial z} = electron\_diff\_current\_is\_opposite\_to\_the\_Electron\_flux\_density;$$

$$(J_p)_{diff} = -qD_p \frac{\partial p}{\partial z} = hole\_diff\_current\_is\_in\_same\_direction\_as\_the\_Hole\_flux\_density;$$

(2.2.9.3)

2.2.10.Life Time of Minority Carriers.

Life Time of minority carriers is the average survival time of minority carriers before it recombines.

In an ideal semiconductor, there are no impurities, no defects and hence no energy states in the band-gap. Therefore once EHP is thermally generated it has very little likelihood to recombine.

2.2.10.1.Two kinds of recombination.

There are two kinds of recombination: in-direct recombination also known as Shockley-Hall-Reed(SHR) recombination and direct recombination known as Auger recombination.

In ideal semi-conductors, the equilibrium between thermal generation and Auger Recombination of EHP decides the equilibrium value and any perturbation of carrier concentration takes a very long time to come to the thermal equilibrium value. In such circumstances the life-time is very large in minutes and seconds.

In real semi-conductors there are energy states introduced in the energy band-gap. Impurities introduce intermediate energy states in the band gap. Imperfections in the crystal structure or crystal defects also introduce intermediate energy states. These intermediate energy states act as the stepping stone for indirect recombination or for SHR recombination as shown in Figure 2.2.41. But the effectiveness of these intermediate energy states as a recombination center depends on the relative position within the band-gap. The middle position within the band-gap makes the intermediate energy states a very effective stepping stone for in-drect recombination. For instance Au. It introduces intermediate energy states which lie in the middle of the Si band-gap hence it is a very effective stepping stone for recombination.Therefore Au is a life-time killer.

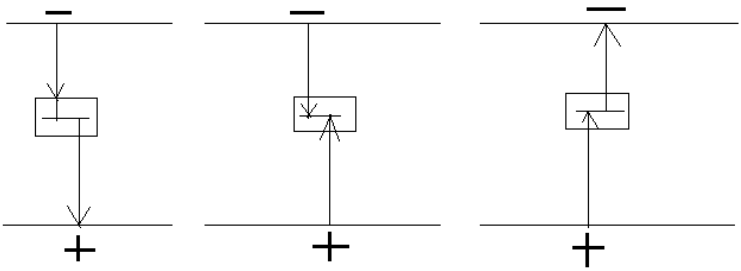


Figure 2.2.41. Three possible mechanisms by which EHP recombination by Shockley-Hall-Reed Recombination can take place using intermediate energy states as the stepping stone. These intermediate energy states are introduced due to the impurities and due to crystal defects

Under very high injection condition, when minority carriers concentration is comparable to majority carriers concentration then Direct Recombination takes place called Auger Recombination as shown in Figure 2.2.42. This situation is encountered in Silicon Control Rectifier when it fires and goes in Saturation mode. It also occurs under large forward bias of PN Junction diode.

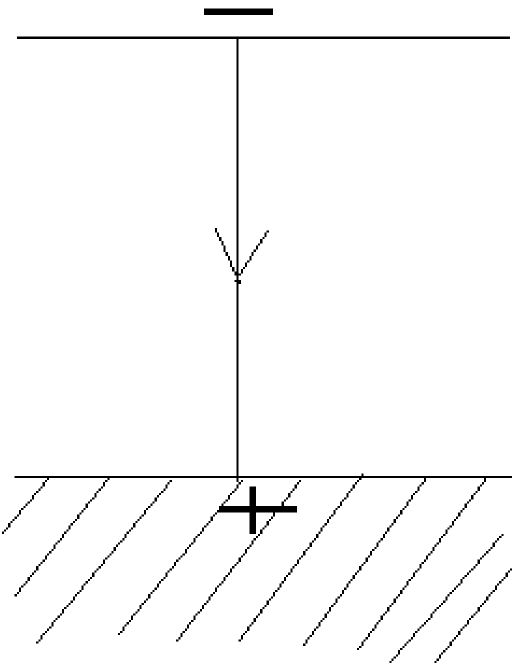


Figure 2.2.42. Direct Recombination or Auger Recombination under very high injection level condition.

2.2.10.2. Decay of Perturbations in equilibrium values of electron and holes. Under normal conditions, conducting electrons and holes are in equilibrium conditions. There are two mechanisms which are perpetually at work within a semiconductor:

Generation of EHP by thermal process or optical process and represented by ‘G’.

Recombination of EHP by indiect recombination process and represented by ‘R’.

Under equilibrium condition: G = R and n and p achieve an equilibrium value which is time – invariant.

There can always be external perturbing factors.

If external perturbing factors persist, n and p will acquire a new equilibrium value.

If external perturbing factor is transitory then perturbation in n and p will be a decaying type transient.

This decaying type transient is a solution of the following differential equation which beautifully embodies the Generation and Recombination process:

$$\frac{dn}{dt} = -\frac{n - n_{\bar{p}}}{\tau_n};\ \frac{dp}{dt} = -\frac{p - p_{\bar{p}}}{\tau_n};$$

(2.2.10.2.1)

Excess carrier decay time constant is always decided by the minority carrier life-time hence in the above Equation both decay rates are controlled by τ_n because in this case electrons are the minority carriers.

In the above equation, LHS gives the rate of change of the carriers and RHS gives the excess carrier divided by carrier Life-Time. If excess carrier is positive, rate of change is negative and if excess carrier is negative then rate of change is positive.

Phyical implication is the following:

1. if the actual carrier is greater than the thermal equilibrium value then the excess carrier decays to equilibrium value by recombination.
2. If the actual carrier is less than the thermal equilibrium value then the deficit carrier builds up to equilibrium value by generation.
3. In both cases the perturbed values exponentially decay and actual values settle to equilibrium values.

Equation 2.2.10.2.1. is a first order ordinary linear differential equation. Its variable is the perturbed values:

$$n - n_p^{-} = n_p^{\wedge}; p - p_p^{-} = p_p^{\wedge}$$

;(2.2.10.2.2)

Rewriting Equation 2.2.10.2.2. in terms of perturbed values:

$$\frac{dn^{\wedge}}{dt} = -\frac{n^{\wedge}}{\tau_n};$$

Or

$$\frac{dn^{\wedge}}{dt} + \frac{n^{\wedge}}{\tau_n} = 0;$$

Therefore the solution of the first order ordinary linear differential equation is:

$$n^{\wedge} = n - n^{-} = Aexp\left(-\frac{t}{\tau_n}\right);$$

A is an arbitrary constant determined by one boundary condition:

At time t=0,

$$n^{\wedge} = n_0 - n^{-} = n_0^{\wedge};$$

Substituting this boundary condition we get: A = nˆ0;

Theefore the solution is rewritten as:

$$n^{\wedge} = n - n^{-} = n_0^{\wedge}exp\left(-\frac{t}{\tau_n}\right);$$

(2.2.10.2.3)

If n > n- then excess carrier recombines exponentially with time constant τ_n;

If n < n- then deficit in concentration is made up by net generation and carrier concentration exponentially increases with time constant τ_n; until the thermal equilibrium value is reached.

Exponential decay of perturbation in thermal equilibrium value of carrier concentration is shown in Figure 2.2.43. for different values of Life-Time.

Life-Time varies from ns to ms depending upon the impurty concentration and crystal imperfections.If impurity concentration and density of imperfection is low then we have lifetime of ‘ms’ interval otherwise we have ‘ns’ interval life-time.

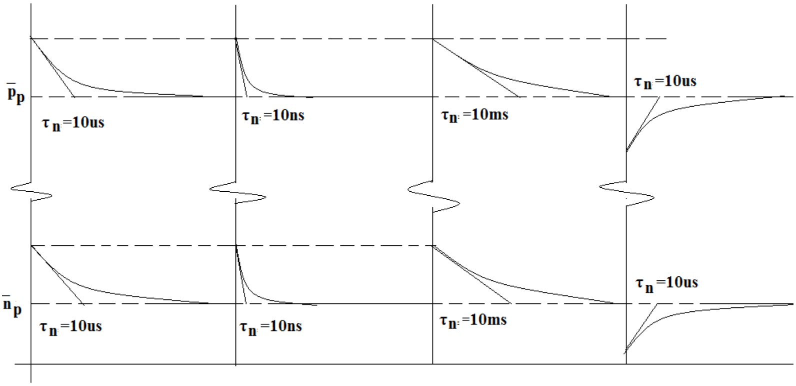


Figure 2.2.43.Exponential decay of the perturbation in the thermal equilibrium value is identical in majority and minority carriers but the exponential decay of the perturbation in all samples have a time -constant equal to life-time of the minority carriers.

2.2.11.Continuity Equation in Silicon Bar optically excited by a continuous beam of light

under steady state condition.

In Figure 2.2.44. N-Type Bar ($N_D = 10^{14}$/cc) is shown which is illuminated by continuous light which has energy packet of 2eV on the left end of the sample.In due course of time the concentration profile of the majority and minority carriers become time-invariant and we say that thermal equilibrium condition has been achieved.

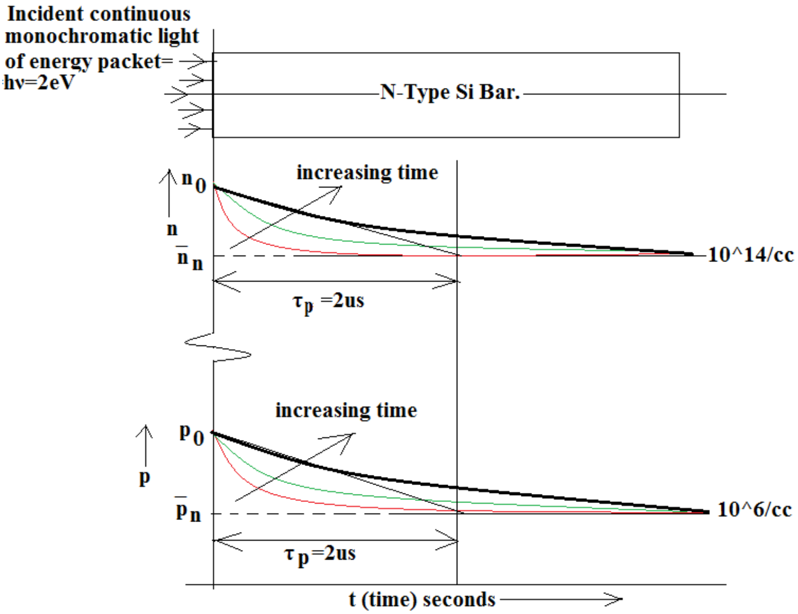


Figure 2.2.44. The concentration profile of majority and minority carriers in an optically perturbed Si-N-Type Bar with increasing time. Red-earliest profile.Green-intermediate profile.Black- steady state thermal equilibrium conc. profile.

Here we have assumed minority carrier hole life-time = 2μs.

Let Optical Generation Rate = G number of EHPs are being opticall generated per second.

The photon of the incident light has an energy packet = 2eV = hν.

Suppose the rate of incidence of energy = 10mW = 10mJ/s.

Since photon is 2eV which exceeds the energy band-gap = 1.12eV hence it is powerful enough to cause photo-excitation. Hence all photons are absorbed by the left surface and equal number of EHPs are optically generated.

The physics of photo-excitation of EHP and recombination of EHP is shown in Figure 2.2.45.

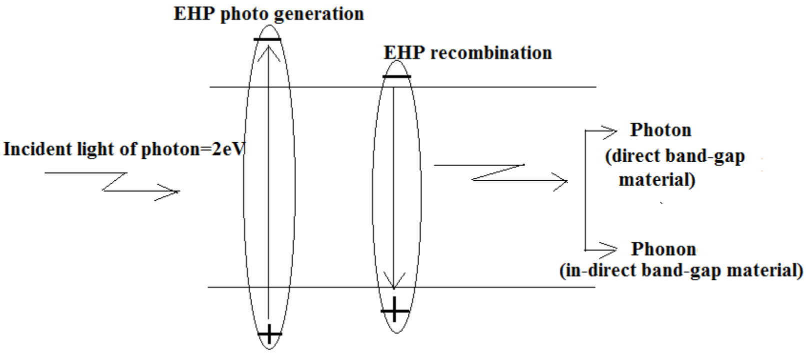

EHP photo generation
EHP recombination
Incident light of photon=2eV
Photon
(direct band-gap material)
Phonon
(in-direct band-gap material)

Figure 2.2.45. Physics of Photo-excitation of EHP and subsequent recombination of EHP. In DIRECT BAND-GAP semiconductor recombination energy generates PHOTON of energy packet=1.12eV and in IN-DIRECT BAND-GAP semiconductor PHONON of energy packet= 1.12eV is generated.

Assuming 1 optically generated EHP for each absorbed incident photon, the optical generation rate G is as follows:

$$G = \frac{10\times10^{-3}\,J/s}{2eV/photon\times1.6\times10^{-19}\,J/eV} = 3.125\times10^{16}\,Photons/s;$$

(2.2.11.1)

Therefore under optical excitation, the left surface experiences a photon boosted generation rate of EHPs $= 3.125\times10^{16}$EHPs/sec.

Let the photons be absorbed within a volume$=\Delta$Vcc.

Therefore optical generation rate -= $g_{op}=10^{13}$EHP/(cc-μs)= 10^{19}EHP/(cc-s).

In DARK condition, through out the bar, thermal generation rate decides the equilibrium concentrations.

In Left End illumination, the photo-generation + thermal generation together decide the equilibrium concentration values on the left end of the Si-Bar.

Let new equilibrium concentration on the left end be p_{n0} and n_{n0}.

Generation-Recombination equation under optical-excitation condition is:

$$g_{op} = 10^{19}\,\frac{EHP}{cc-s} = R = \frac{p_{n0}-\overline{p}_{n}}{\tau_{p}} = \frac{n_{n0}-\overline{n}_{n}}{\tau_{p}};$$

$$Therefore\_\overline{\overline{p}}_{0} = p_{n0}-\overline{p}_{n}\_and\_\overline{\overline{n}}_{0} = n_{n0}-\overline{n}_{n};$$

$$\overline{p}_{0} = \overline{n}_{0} = 10^{19}\left(\frac{EHP}{cc-s}\right)\times 2\times 10^{-6}\,s = 2\times 10^{13}\left(\frac{number}{cc}\right);$$

(2.2.11.2)

Now Law of Mass Action does not hold good because of external illumination.

Therefore $n_{n0} \times p_{n0} \neq n_{i2}$;

Some time after illumination, carrier concentration assumes a steady state value through out the Si-Bar along the z-axis.

This steady state concentration profile has to be determined.

Let us consider holes only.

Refer to Figure 2.2.46.

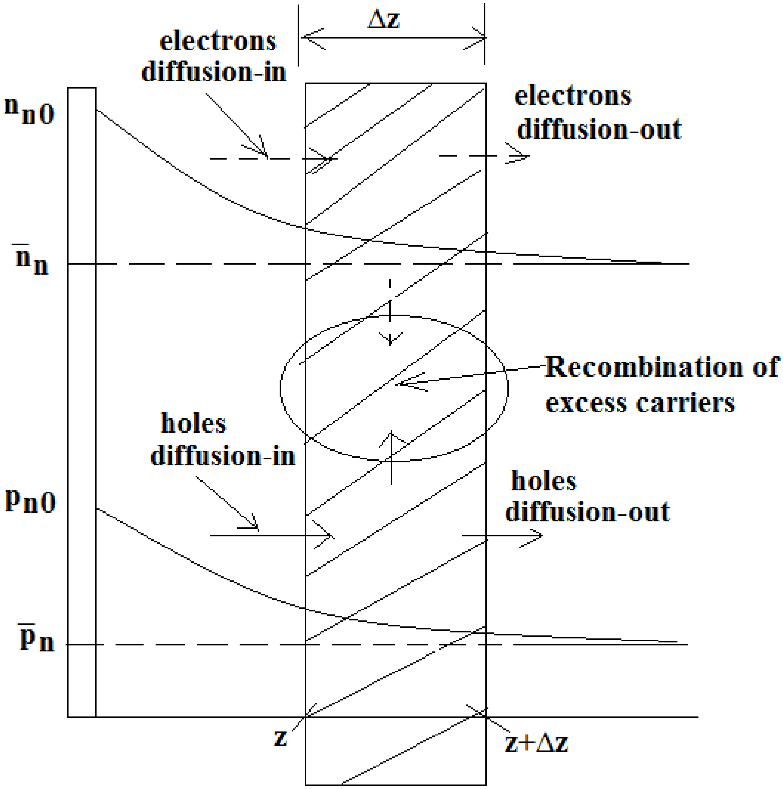


Figure 2.2.46.Setting up the Continuity Equation by looking at the physical processes in elemental volume A×Δz.

$$Right\_ward\_hole\_flux\_density\_at\_z:$$

$$j_p(z) = -D_p \frac{\partial p_n(z)}{\partial z} \Big| z;$$

$$Right\_ward\_hole\_flux\_density\_at\_z+\Delta z:$$

$$j_p(z+\Delta z) = -D_p \frac{\partial p_n(z)}{\partial z} \Big| (z+\Delta z);$$

(2.2.11.3)

$$\textit{Therefore\_inward\_hole\_diffusion} = \textit{hole\_recombining\_in\_elemental\_volume} + \textit{outward\_hole\_diffusion};$$

$$\textit{Therefore\_} j_p \Big| z = \frac{p_n - \overline{p_n}}{\tau_p} \times \Delta z + j_p \big| z + \Delta z;$$

$$\textit{Therefore\_} \frac{j_p|z - j_p|z + \Delta z}{\Delta z} = \frac{p_n - \overline{p_n}}{\tau_p};\underline{\qquad\qquad\qquad\qquad}(2.2.11.4)$$

$$\textit{By\_definition\_First\_Derivative\_of\_}\text{diffusion\_flux\_density}:$$

$$\frac{\partial j_p}{\partial z} = \frac{j_p|z + \Delta z - j_p|z}{\Delta z};$$

$$\textit{Theefore\_Equation}2.2.11.4\textit{\_is\_rewritten\_as}:$$

$$-\frac{\partial j_p}{\partial z} = \frac{p_n - \overline{p_n}}{\tau_p};\underline{\qquad\qquad\qquad\qquad\qquad\qquad\qquad}(2.2.11.5)$$

$$\textit{But\_} j_p = -D_p \frac{\partial p_n(z)}{\partial z};$$

$$\text{Therefore\_Equation 2.2.11.5 \_is\_rewritten\_as}:$$

$$\mathrm{D}_p \frac{\partial^2 p_n(z)}{\partial z^2} = \frac{p_n - \overline{p_n}}{\tau_p};$$

$$\textit{Therefore\_} \frac{\partial^2 p_n(z)}{\partial z^2} = \frac{p_n - \overline{p_n}}{\tau_p D_p};$$

$$\text{Let\_}\tau_p D_p = L_p = \text{diffusion\_length};$$

$$\textit{Therefore\_} \frac{\partial^2 p_n(z)}{\partial z^2} = \frac{p_n - \overline{p_n}}{L_p^2};$$

$$\text{Express\_this\_in\_excess\_carrier\_concentration}:$$

$$\text{Therefore\_} \frac{\partial^2 \overline{p_n}}{\partial z^2} = \frac{\overline{p_n}}{L_p^2};$$

$$\textit{This\_is\_a\_s tan dard\_sec ond\_order\_}\text{ordinary\_}\textit{linear\_differential\_equation}$$

$$\textit{with\_s tan dard\_solution}:$$

$$\overline{p_n} = A\exp(-\frac{z}{L_p}) + B\exp(\frac{z}{L_p});$$

$$\text{By\_setting\_the\_boundary\_conditions\_namely}:$$

$$\textit{At\_} z = 0, \overline{p_n} = \overline{p_{n0}}\textit{\_and\_at\_} z = \infty, \overline{p_n} = 0;$$

$$\textit{We\_get}:$$

$$\overline{\overline{p_n}} = \overline{\overline{p_{n0}}}\exp(-\frac{z}{L_p});$$

$$\text{Similarly}:$$

$$\overline{n_n} = \overline{n_{n0}}\exp(-\frac{z}{L_p});$$

(2.2.11.6)

The typical diffusion length in this particular example is:

$L_p^2 = \tau_p D_p = 2 \times 10^{-6} \times 12.5 cm^2/s = 25 \times 10^{-6} cm^2$;

Therefore diffusion length $= L_p = 5 \times 10^{-3} cm = 0.5 \mu m$.

In Figure 2.2.47. steady state carrier concentration profiles are given with one end of the Si-Bar photo-excited and for the case Length of the Bar 'L' > > Lp. This is a wide-bulk case.

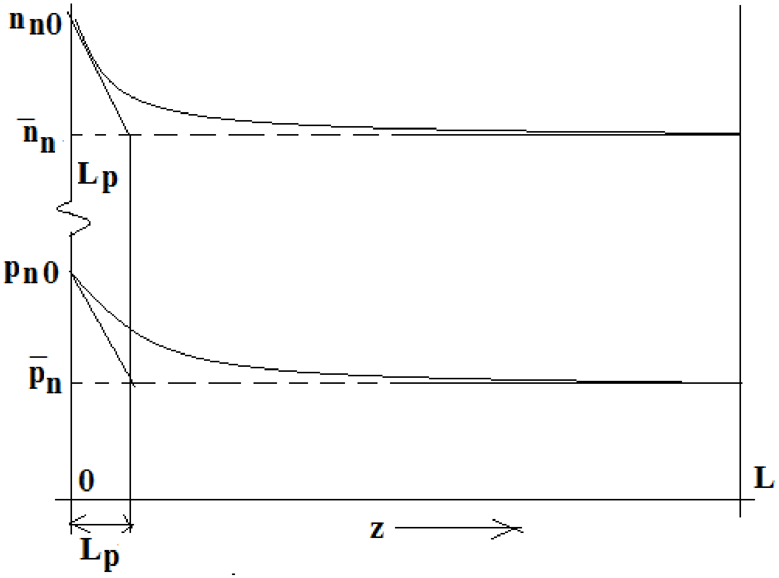


Figure 2.2.47. The steady state carrier concentration profile in a wide-bulk Si-Bar with left hand end continuously photo-excited.

In Si-Bar, when we have L ~ Lp or L < < Lp then we have narrow-bulk case and steady state carrier concentration profile is linear as shown in Figure 2.2.48.

The linearity of the concentration profile is due to the fact that in a narrow bulk thee is hardly any recombination hence diffusion current is constant along the z-axis and hence carrier concentration of the carrier is also constant. This leads to a linear slope within the bulk for both majority and minoreity carriers.

In our example, in case of narrow-bulk Si-Bar, the far end steady state concentration values depend upon surface recombination velocity. In our case we have assumed infinite surface recombination velocity hence:

$$p_n(L)=\overline{p_n}\_and\_n_n(L)=\overline{n_n};$$

(2.2.11.7)

The case for finite surface recombination velocity will be dealt in the Appendix of this Chapter.

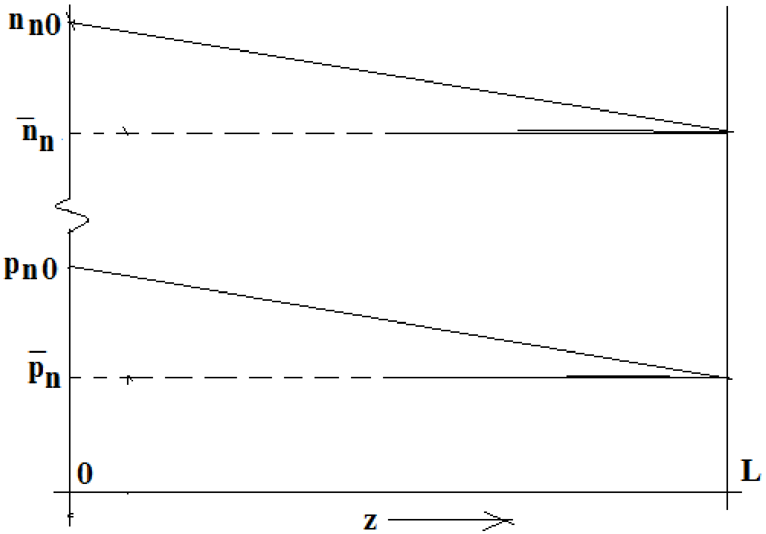


Figure 2.2.48. The steady state carrier concentration profile in narrowbulk Si-Bar with left hand end continuously photo-excited.

Tutorial Sheet of Chapter 2.Semiconductor Physics. This gives the tutorial sheet on Chapter 2- Semiconductor Physics.

Tutorial Sheet of Chapter 2.Semiconductor Physics.

Question 1. (a)Determine the Drift Velocity of electron and hole in the given intrinsic sample in Figure-Tutorial 2.1. with 10V applied across the junction?[vdrift(n)=5800cm/s,vdrift(p)=1800cm/s]Assume mobility given for intrinsic sample.

(b)Determine the conductivity of the following intrinsic sample?[$\sigma=3*10^{-6}$S/cm]

(b) Determine the current flowing though the intrinsic sample?[0.72microA]

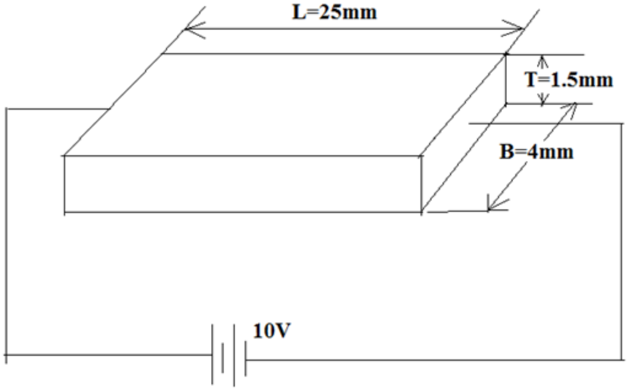


Figure-Tutorial2.1 Intrinsic Sample with a longitudinal Electric Field.

Question 2. The energy-band diagram is given in Figure-Tutorial2.2.

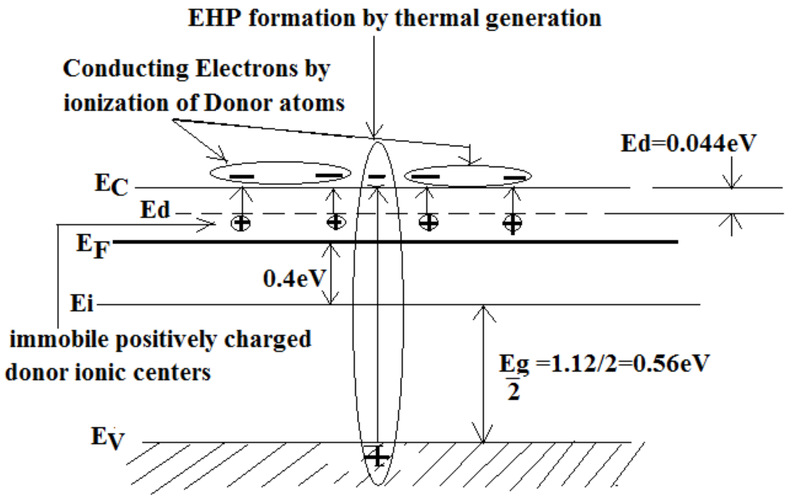

EHP formation by thermal generation
Conducting Electrons by ionization of Donor atoms
Ed=0.044eV
EC
Ed
EF
0.4eV
Ei
immobile positively charged donor ionic centers
Eg/2 =1.12/2=0.56eV
EV

Figure-Tutorial2.2. Energy-Band Diagram of N-Type Material.

1. Calculate the probability of occupancy at Ed according to Fermi-Dirac Statistics? [P(Ed) = 1.41%]
2. Calculate the % of ionization of Donor Atoms ? [99%]
3. If N_C = effective density of states at E_C = 3.01×10^{19}/cc then calculate n_{n-} ? [6.3917×10^{16}/cc]
4. Calculate p_{n-} if N_V = effective density of states at E_V = 1.04×10^{19}/cc ? [2773.81]
5. Calculate n_i^2 ? [$(1.15 \times 10^{10})^2$]
6. Calculate N_D [6.3917×10^{16}/cc]

Question 3. In a P-Type material, E_F-E_V = 0.1 eV and E_a-E_V = 0.045eV.

Calculate the % of ionization of Acceptor Atoms ? [89%]

Question 4. There is a compensated P-Type material at 300K. Minority carrier concentration is measured to be $n_{p-} = 10^{6}$/cc. Acceptor Dopent Concentration $= 10^{15}$/cc.

1. Determine majority carrier concentration? [10^{14}/cc].
2. Determine N_d ? [9×10^{14}/cc].
3. Determine the total scattering centers. [19×10^{14}/cc].

Question 5. In Copper, resistivity $= \rho = 1.67 \times 10^{-8}$ Ω-m and conductivity $= \sigma = 6 \times 10^{7}$ S/m $= q\mu n$ and $n = 8.5 \times 10^{22}$/m^{3}. Determine electron mobility ? [40 cm^{2}/(V-s)]

Question 6. In a 3 m long Cu Wire R = 0.03Ω and I = 15A through the wire.

1. Determine V_{drop} across the wire ? [0.45V]
2. Determine E along the z-axis ? [0.15V/m]
3. Determine drift velocity of electron if $\mu =$ 40 cm^{2}/(V-s)? [0.6×10^{-3}m/s]
4. Determine the Fermi-velocity of conducting electron ? [10^{6}m/s]
5. Determine the mean free time between two scatterings also defined as relaxation time ? [2.27×10^{-14}s]
6. Determine the mean free path ? [275A°].

Compare this with lattice constant a $=3.6$A°.

Question 7. Typical N-Type resistivity $\rho_s = 1\Omega$-cm.Typical current density 100A/cm^2. mobility of electrons$=1000$cm^2/V-s

1. Calculate E ? [100V/cm]
2. Calculate drift velocity ? [1.0×10^5cm/s]

Question 8. In a P-Type Sample , E$=10^3$V/cm is applied. Determine drift velocity, mean free time and mean free path ? Assume $\mu_p = 470$ cm^2/(V-s).

[Drift velocity $= 4.7\times10^5$cm/s which is much lower than thermal velocity of holes$=2.2\times10^7$cm/s; $\tau = 0.1$ps, Mean Free Path $=$ 220A°]

Question 9. The effective mass of electron and hole in semi-conductor is different from the mass in free space . Why?

The electrons and holes in a crystal interact with the periodic Coulombic Field in the Crystal. They surf over the periodic potential variation of the crystal lattice in the process developing roller coaster effect which leads to drastic reduction in effective mass.

	Si	**Ge**	**GaAs**	**InAs**	**AlAs**
m_n/m_0	0.26	0.12	0.068	0.023	2.0
m_p/m_0	0.39	0.30	0.50	0.30	0.3

m_n/m_0 = effective mass of electron/free space mass of electron

m_p/m_0 = effective mass of hole/free space mass of hole.

In Silicon, using the effective mass of electron and holes the thermal velocity = 2.5×10^{7}cm/s. Scatter Limit Velocities of electron and hole are of the same order.

SSPD_Chapter 3_Solid State Diode Physics.
Section 3. covers the Diode Physics and its theoretical formulation under Ideal Condition and under Real Condition.

SSPD_Chapter 3_Solid State Diode Physics.

Today we are living in knowledge intensive society which is going through the Third Information Revolution. In this Information Age we are critically dependent upon Solid State Devices for generating, processing and transmitting information and information is nothing but signal. Hence Solid State Devices are performing the following signal processing:

1. Reception/detection of signal. This signal can be very weak and heavily corrupted by noise and interferences.
2. Amplification of Signal through Signal Amplifiers.
3. Signal Processing for restoring the integrity of the signal.
4. Information carried by the signal needs to be stored in the memory. There are compact discs and magnetic tapes.Gradually these are being supplanted by Solid State Memories. They are cheap and are becoming faster.
5. Through Solid State Sensors, environmental parameters are being sensed and converted into electrical/optical signals which are

subsequently processed and utilized for different applications.

6. Solid State Displays and Flat Panel Displays utilizing Solid State Devices are increasingly taking the place of the conventional displays.

One of the Solid State Devices which has been in use since early 20th century is PN Junction Diode. It has been used for envelope detection of Amplitude Modulated Signals in Radio Broadcasts.

Subsequently it has been used in power conditioning circuits for generating DC Voltages from AC Voltages.

It has found applications in Frequency Modulation Circuits as Varactor Diodes.

It has found applications in MicroWave Generation and Amplification as Esaki or Tunnel Diodes.

It has found applications as Schottky Diodes in Scottky Clamped TTL Circuits for speeding up the switching speed of TTL Gates.

PN Junction can be modified to behave like ohmic contacts instead of rectifying contacts.

PN Juntions are extensively used as Photo-Voltaic Cells or as Solar Cells.

3.1. Fabrication of PN Junction Diode.

A PN Junction diode is fabricated by diffusing N-type dopent(Phosphorous) into P-Type Wafer with Boron Doping at an uniform concentration of 10^{15}/cc corresponding to a resistivity of 10 Ω-cm. The diffusion is carried out in Diffusion Furnace at a high temperature of 1000°C to 1300°C with a tolerance of ±0.5°C. The diffusion furnace is shown in Figure 6.14 in Chapter 6.

There can be two kinds of diffusion shown in Figure 3.1:

Pre-depostion- It is a short period constant-source diffusion with a shallow junction. The Doping Profile is Complementary Error Function and the Junction is a one-sided step-junction or abrupt junction.

Limited Source Diffusion- It is a long drive-in diffusion [erfc(z)] with a deep junction. The doping profile is Gaussian and the junction is linearly graded junction as shown in Figure 3.1.

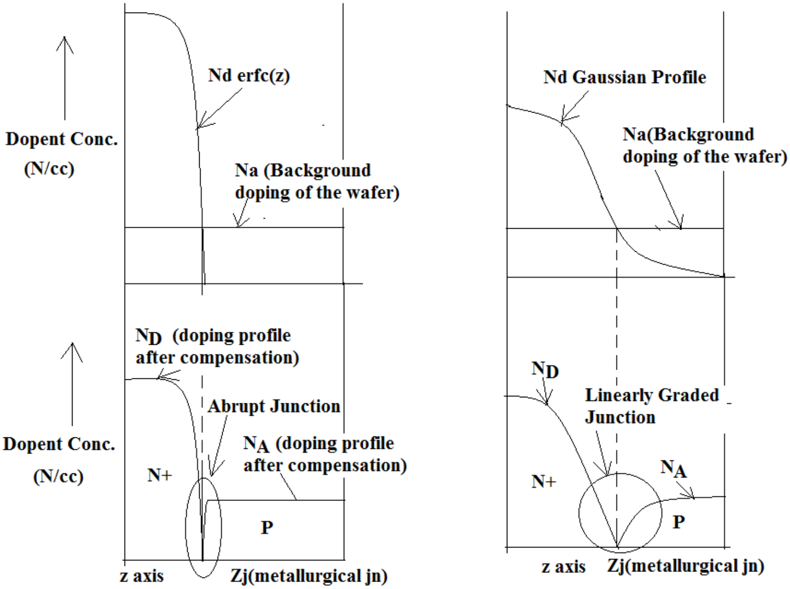


Figure 3.1. Abrupt Junction and Linearly-Graded Junction Diode.

.Both these two kinds of junctions have their unique electrical parameters which we will see in the subsequent section.

3.2.Physics of PN Junction Diode.

Whenever two dis-similar materials are brought in contact, a contact potential is developed. This property is used in Thermo-couple.

Similarly when P-Type and N-Type is brought in contact, a built-in barrier potential (Φ_{BO}) develops at the metallurgical junction of P-Type and N-Type semiconductor. As shown in Figure 3.2., diffusion of the majority carriers leads to the development of a dipole layer at the junction.

This dipole layer is called space-charge layer, transition layer or depletion layer. The development of dipole layer prevents further diffusion of majority carriers.

This built-in barrier potential is always an up-hill for the majority carriers and is always a down-hill for minority carriers.

The diffusion of majority carriers from high concentration to low concentration is diffusion current and drift of minority carriers down the potential hill is drift current.

At this point a detailed charge balance is achieved as follows:

$$I_{n_{diff}}(from\ N-side\ to\ P-side)=I_{n_{drift}}(from\ P-side\ to\ N-side)$$

3.2.1

$$I_{p_{diff}}(from\ P-side\ to\ N-side)=I_{p_{drift}}(from\ N-side\ to\ P-side)$$

3.2.2

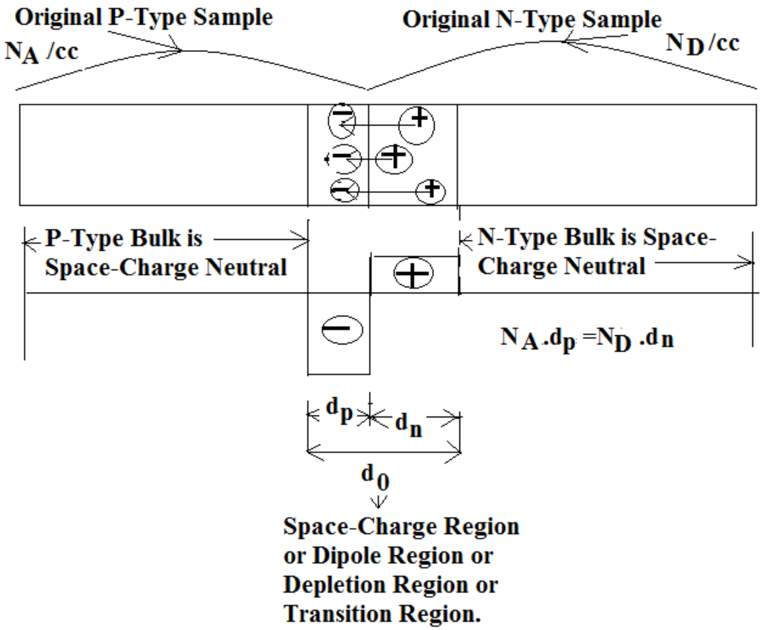

Original P-Type Sample
N_A /cc
Original N-Type Sample
N_D/cc
P-Type Bulk is Space-Charge Neutral
N-Type Bulk is Space-Charge Neutral
$N_A.d_p = N_D.d_n$
d_p
d_n
d_0
Space-Charge Region or Dipole Region or Depletion Region or Transition Region.

Figure 3.2. The development of dipole layer at the metallurgical junction due to the depltion of majority carriers and exposure of the immobile ionized dopent atoms on the two sides of the junction.

A more detailed picture of the depletion region is given in Figure 3.3 under zero-bias condition. This gives the space-charge profile, electric-field profile and the potential profile along the longitudinal axis of the PN Junction diode.

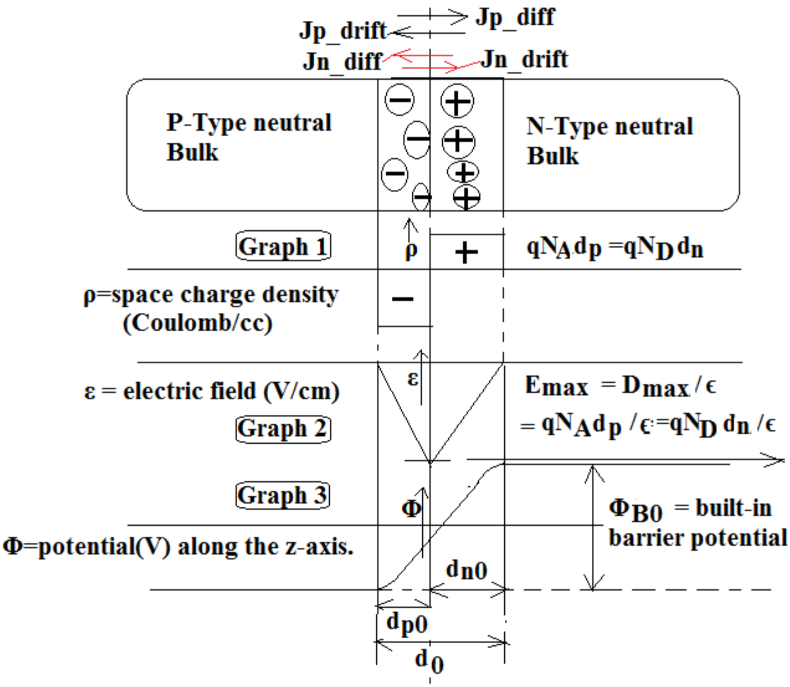

Jp_drift
Jp_diff
Jn_diff
Jn_drift
P-Type neutral Bulk
N-Type neutral Bulk
Graph 1
ρ
$qN_A d_p = qN_D d_n$
ρ=space charge density (Coulomb/cc)
ε = electric field (V/cm)
ε
Graph 2
Emax = Dmax / ϵ
$= qN_A d_p / \epsilon = qN_D d_n / \epsilon$
Graph 3
Φ
Φ_{B0} = built-in barrier potential
Φ=potential(V) along the z-axis.
dn0
dp0
d_0

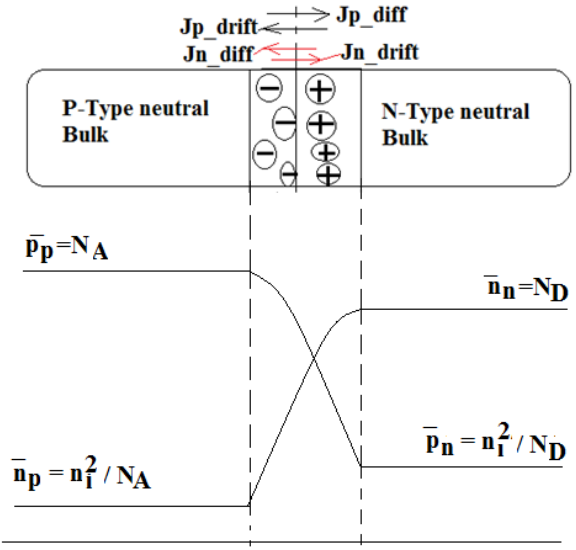
Figure 3.3.Space-Charge Profile, Electric Field Profile and Potential Profile along the z-axis of the PN-Jn. Diode under zero-bias.

Under ZERO-BIAS condition,as seen in Figure 3.3. **Holes,** majority carriers in P-Bulk, are diffusing from P-Region to N-Region. **Holes,** minority carriers in N-Bulk, which survive to reach the edge of the depletion region just flow down-hill across the potential gradient as shown in Graph 3 in Figure 3.3. φ

As also seen in Figure 3.3. **Electrons,** majority carriers in N-Bulk, are diffusing from N-Region to P-Region. **Electrons,** minority carriers in P-Bulk, which survive to reach the edge of the depletion region just flow down-hill across the potential gradient as shown in Graph 3 in Figure 3.3.

Under zero-bias, hole diffusion is exactly balanced by hole drift and electron diffusion is exactly balanced by electron drift. Hence by detailed charge balance net current through the Diode under zero-bias is zero.

Graph 1 gives the Space Charge Density (ρ) C/cc. In the bulk region there is no space charge. The positive and negative are exactly balanced hence net charge is zero. But in depletion region d_0, there is uniform negative space charge density on the P-Side extending over d_{p0} cm and there is uniform positive space charge density on N-Side extending over d_{n0} cm. The negative space-charge is due to negatively ionized Boron Atoms on P-Side and positive space charge is due to positively ionized Phosphorous Atoms on N-Side.Since I am discussing One-Sided Step Junction or Abrupt Junction hence doping density is uniform as seen in Figure 3.1. This is the reason why Space-Charge Density in depletion region is uniform on both sides though of opposite polarities.



$\bar{p}_p$ = thermal equilbrium hole density (majority carriers) on P-Side

$\bar{n}_p$ = thermal equilbrium electron density (minority carriers)on P-Side

Similarly $\bar{n}_n$(majority carriers on N-Side) and $\bar{p}_n$ (minority carriers on N-Side) are defined.

Figure 3.3.A. Carrier Concentration Profile along the z-axis of PN Jn diode.

On P-Side, Acceptor Dopent Density N_A is uniform from $-d_{p0}$ to zero. On N-side , Donor Dopent Density is uniform from 0 to d_{n0}.

Since depletion region is a dipole layer with equal positive and negative charges on the two poles hence:

$$qN_Ad_{p0} = qN_Dd_{n0} \qquad 3.2.3$$

Equation (3.2.3) tells us that *depletion layer is deeper on the lightly doped side and shallower on heavily doped side.*

Graph 2 gives the Electric Field Profile. Electric Field(E) is Electric Flux Density (D) divided by permittivity($\epsilon = \epsilon_0\epsilon_r$) where D is the number of Flux passing through cm^2 cross-sectional area. In Rationalized MKS units it is assumed that 1 Coulomb of Charge gives rise to 1 Coulomb of electric Flux. Hence qN_Dd_{n0} is the total positive charge within a volume of $1cm^2$ cross-sectional area and length d_{n0} cm. Equivalently qN_Ad_{p0} is the total negative charge within a volume of $1cm^2$ cross-sectional area and length d_{p0} cm. Electric Flux originates on +ve Charge and terminates on –ve charge. Therefore Flux Density D is maximum at metallurgical junction $z = 0$ and negatively directed.

Therefore at $z = 0$, $D_{max} = -qN_Dd_{n0} = -qN_Ad_{p0}$ and on the two sides D linearly decreases to zero at $z = -d_{p0}$ and at $z = d_{n0}$. Hence E is maximum at $z = 0$ and falls linearly to zero at $z = -d_{p0}$ and at $z = d_{n0}$ as shown in Graph 2. From this we conclude that

$$E_{max} = -\frac{qN_Ad_{p0}}{\epsilon} = -\frac{qN_Dd_{n0}}{\epsilon} \qquad 3.2.4.$$

Since Electric Field E is the negative of potential gradient hence potential is the area under the Electric Field profile:

$$\varphi_{BO} = built - in\ barrier\ potential = area\ of\ the\ triangle = \frac{d_0 \times E_{max}}{2} \qquad 3.2.5.$$

Therefore:

$$\varphi_{BO} = \frac{d_0}{2} \times \frac{qN_A d_{p0}}{\in} = \frac{d_0}{2} \times \frac{qN_D d_{n0}}{\in} \qquad 3.2.6.$$

Therefore:

$$d_0 d_{p0} = \frac{2 \in}{q} \times \frac{\varphi_{BO}}{N_A} \qquad 3.2.7$$

Similarly

$$d_0 d_{n0} = \frac{2 \in}{q} \times \frac{\varphi_{BO}}{N_D} \qquad 3.2.7.$$

Adding Equation (3.2.7) and (3.2.8) we get:

$$d_0^2 = \frac{2 \in \varphi_{BO}}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right) \qquad 3.2.8.$$

Square rooting Equation (3.2.8):

$$d_0 = \sqrt{\frac{2 \in \varphi_{BO}}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right)} = \sqrt{\frac{2 \in}{q} \left(\frac{1}{N_A} + \frac{1}{N_D} \right)} \times \sqrt{\varphi_{BO}} \qquad 3.2.9.$$

This equation tells us that if we have one sided step junction then depletion layer will lie on lightly doped region and will be determined entirely by the lighter doping density.

Equation (3.2.9) tells us that in abrupt junctions the depletion width ‘d’ is directly proportional to the square root of the barrier potential.

In linearly graded junctions ‘d’ is directly proportional to the cube root of the barrier potential. This will be shown in the advanced version of Diode Physics.

As seen in Figure 3.3.A, there is a carrier concentration gradient in the depletion region which causes the diffusion current across the depletion region. As seen in Graph 3 of Figure 3.3. there is a potential gradient also in the depletion region. This leads to the drift of minority carriers.Drift and Diffusion together lead to detailed charge balance leading to zero current across the depletion under zero-bias condition.

3.2.1. Energy Band Diagram interpretation of Built-in Barrier Potential.

When P-Type Semiconductor and N-Type Semiconductor are alloyed together or PN-Abrupt Junction formed by diffusion then under no bias condition and under thermal equilibrium condition, Fermi-Level of P-Type and that of N-Type align themselves as shown in Figure 3.4 since Fermi-Level is a function of temperature and the two samples are at the same temperature when alloyed together. In the process there is band-bending and built-in potential gradient is created at the junction.

In Chapter 3 Advanced version of diode, it will be shown that:

$$\varphi_{B0} = \frac{kT}{q} ln \frac{N_A N_D}{n_i^2} \qquad 3.2.10.$$

From Equation 2.2.6.2.7:

$$n_i^2 = N_C N_V Exp\left[-\frac{E_G}{kT}\right] \qquad 3.2.11$$

Substituting Equation (3.2.11) in (3.2.10) we get:

$$\varphi_{BO} = V_{Th} \times \ln\left[\frac{N_A N_D}{N_C N_V}\right] + V_G \qquad 3.2.12$$

In Equation (3.2.12), V_{Th} is the thermal voltage kT/q= 0.026V at 300K, V_G= band gap voltage of the material and for Si it is 1.12V and N_C and N_V are equivalent density of states= 10^{19}/cc (for Si) and if N_A= N_D=10^{16}/cc then built-in barrier potential is of 0.8V. .

From Equation (3.2.12), it is evident that Built-in Barrier Potential is of the order of the Band-Gap of the material. Large Band-Gap material such as GaAs will have a large Built-in Barrier Potential and Narrow Band-Gap material such as Ge has a low Built-in Barrier Potential. This is the reason why Ge has a cut-in voltage of 0V, Si has 0.5V and Ga As has 1V.

Equation (3.2.12) also tells us that when doping approaches equivalent density of states, φ_{BO}(Built-in

Barrier Potential) approaches V_G(Band-Gap Volt) and E_F(Fermi-Level) enters the conduction band and valence band in N-Type and P-Type semiconductor respectively as shown in Figure (3.5). The semiconductor becomes de-generate and Maxwell-Boltzmann Approzimation does not hold good and we must apply Fermi-Dirac Distribution.

Degenerate Semiconductor is referred to as semi-metal. Al(Aluminum) is P-Type material and when deposited over lightly doped N-Type material, PN junction is a rectifying contact and is referred to as Schottky Diode, hot electron diode.It has a very high switching speed.This technique is used in Scottky-clamped TTL gate,74 series. But if Al contact is deposited over N+ layer then we get ohmic contact. In vertical IC npn transistor, this technique is used for making contact to the Collector of the BJT.

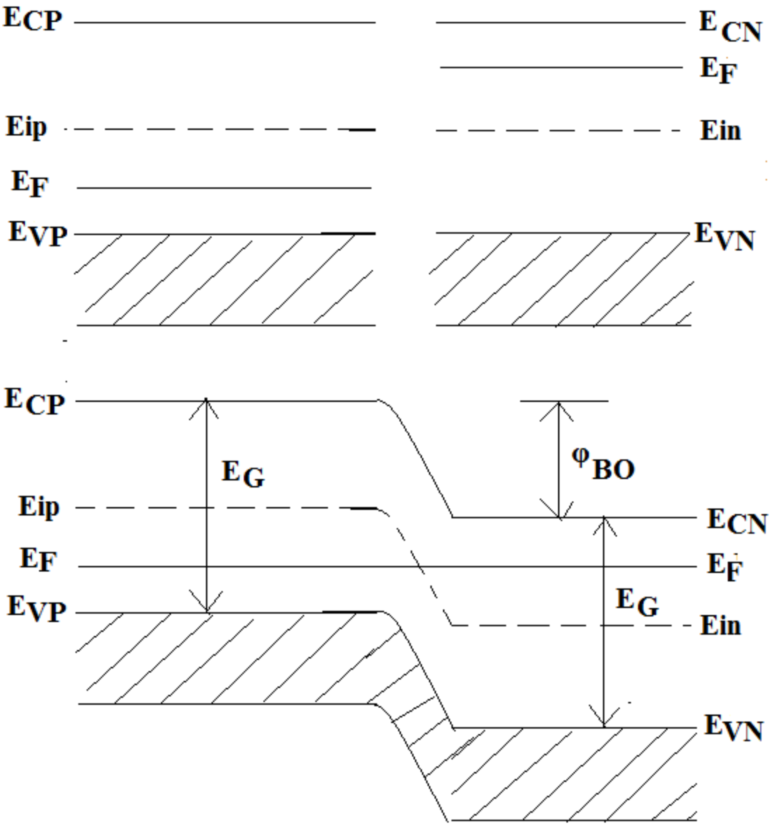


Figure 3.4.Energy Band Diagram of PN Junction diode under zero-bias.

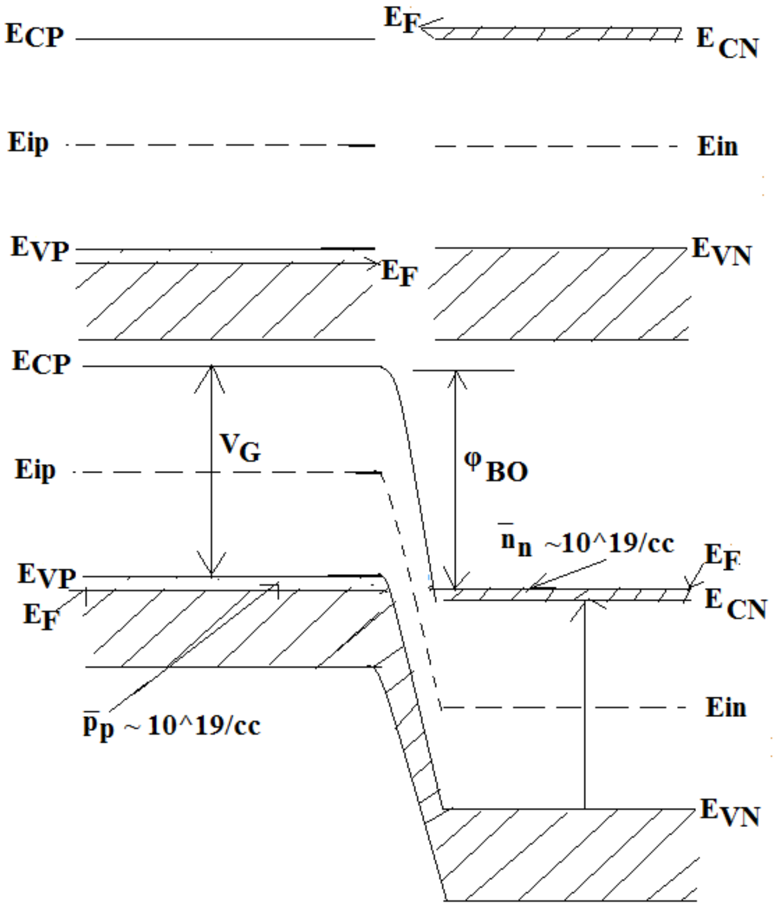


Figure 3.5. Energy Band Diagram of Degenerate Semiconductor

3.2.2.PN Jn Diode under forward bias and theoretical deduction of Shockley Equation or Ideal Diode Equation..

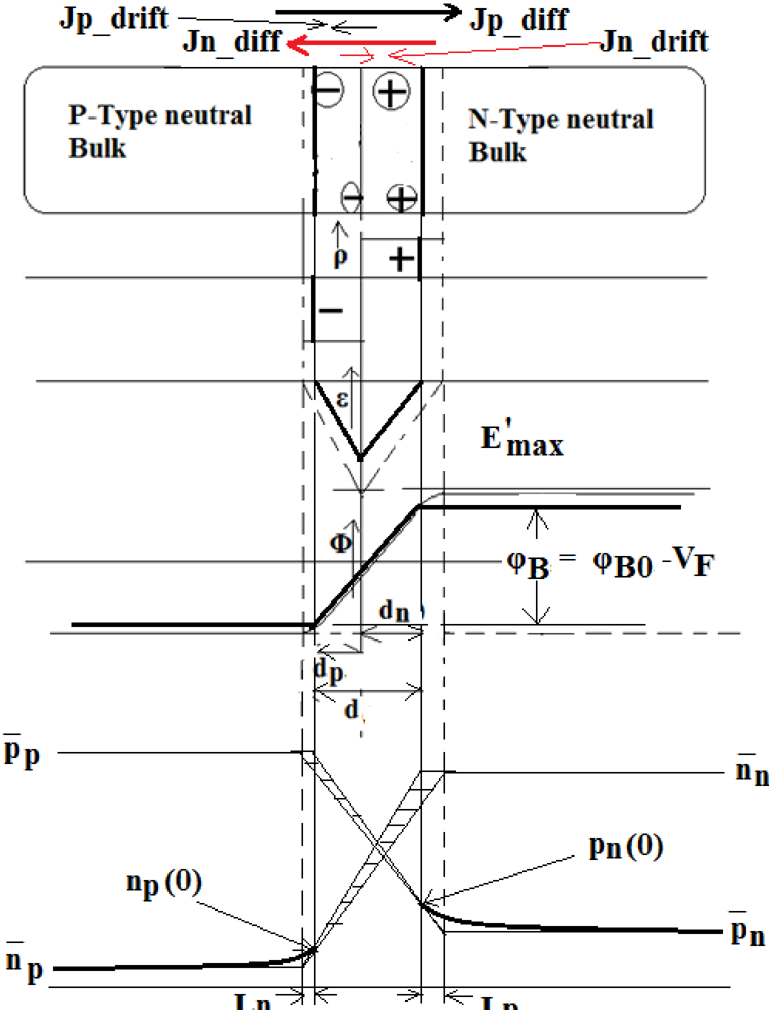


Figure3.6.PN Jn Diode under Forward Bias V_F

Under forward bias, depletion layer shrinks, E_{max} reduces and built-in barrier potential reduces by V_F and majority carriers easily climb over the up-hill potential gradient and get injected to the other side strongly perturbing the minority carrier there-on.

If the forward bias is more than 0.5V which is the cut-in voltage for Si-Diode, PN Jn Diode turns on and there is exponential increase in the diffusion current whereas drift current remains unaltered. Drift current remains unaltered because it depends on the minority carriers falling down the potential gradient in the depletion width. Since minority carriers are in very short supply hence their drift is unaffected by forward bias.But the barrier potential which was of 0.8V drastically reduces for forward voltage greater than 0.5V hence majority carriers easily cross the up-hill potential gradient within the depletion and lead to a decade increase in the forward current for every 60mV increase in the forward bias which we will see in the next section.

Forward Current Density flow , its components and carrier concentration profiles under forward bias condition are shown in Figure 3.7.

As can be seen in Figure 3.7, majority carriers are injected to the oher side strongly perturbing the minority carrier from thermal equilibrium value to a new elevated value under forward bias condition. For example on N-Side:

$$p_n^- \text{ gets elevated to } p_n(0) = p_n^- Exp\left[\frac{V_D}{V_{Th}}\right] \text{obtained from Boltzmann relation}$$

$$Similarly\ n_p^- \text{ gets elevated to } n_p(0) = n_p^- Exp\left[\frac{V_D}{V_{Th}}\right].$$

These new perturbations $p_n(0)$ at the edge of the depletion layer on N-side and $n_p(0)$ at the edge of the depletion layer on P-side exponentially decay to thermal equilibrium values on their respective sides with length constant of L_p and L_n as shown in Figure 3.7.

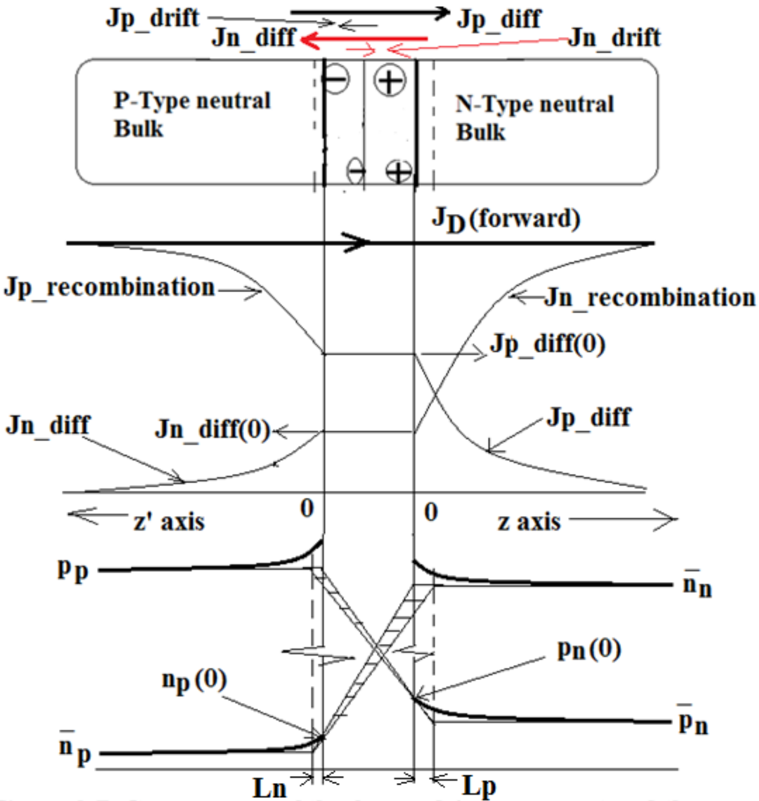


Figure 3.7. Components of the forward diode current and the carrier concentration profile within the diode along longitudinal axis.

As already discussed, forward bias lowers the built-in barrier potential hence though up-hill for majority carriers, majority carriers cross it with ease. Hence diffusion current component carried by majority carriers increases mani-folds and diode is turned on. As seen in Figure 3.7. Depletion layer shrinks and Majority carriers are injected into the other side across the depletion layer. Hence minority carrier profile is perturbed near the edge of the depletion width on both sides of the junction.The perturbation of minority carrier profile leads to identical perturbation in the majority carrier profile for maintaining the space-charge neutrality. The perturbed profile in the minority carrier leads to the diffusion of minority carriers away from the junction to the ohmic contact.As minority carriers diffuse away they recombine with the majority carriers along the way. Hence majority cariers rush in as recombination current from the ohmic contact towards the depletion layer.

For the continuity of the current flow, at every section of the bulk:

$$J_{minority\_diff} + J_{maj\_recombination} = J_D(forward) \tag{3.2.2.1}$$

Within the depletion layer as shown in Figure 3.7, electron and hole currents are constant and there is no recombination within the depletion layer.

Deduction of Shockley Equation or Ideal Diode Equation.(*Optional*)

William Shockley, who invented BJT along with Bardeen and Braittain, made two important assumptions while deriving Ideal Diode Equation:

Assumption 1: There is no recombination within the depletion region and both electron and hole current are constant while crossing the depletion region;

Assumption 2:The two bulk regions are quasi-neutral and the forward applied voltage drops across the depletion region causing the built-in barrier potential to drop from φ_{BO} to (φ_{BO}-V_F). During reverse bias, the built-in barrier potential increases from φ_{BO} to ($\varphi_{BO}+V_R$).

The total diode current J_D(forward)referring to Figure 3.7:

$$J_D(forward) = J_{n\_diff}(z' = 0) + J_{p_{diff}}(z = 0) \qquad 3.2.2.2.$$

$$where\ J_{n_{diff}}(z' = 0) = q D_n \frac{n_p(z' = 0) - n_p^-}{L_n} \qquad 3.2.2.3.$$

By Boltzmann Relationship:

$$n_p(z' = 0) = n_p^- Exp\left(\frac{V_F}{V_{Th}}\right); p_n(z = 0) = p_n^- Exp\left(\frac{V_F}{V_{Th}}\right); \qquad 3.2.2.4$$

Substituting Equation (3.2.2.4) in Equation(3.2.2.3):

$$J_D(forward) = \left[q\frac{D_n}{L_n}n_p^- + q\frac{D_p}{L_p}p_n^-\right]\left[Exp\left(\frac{V_F}{V_{Th}}\right) - 1\right] \qquad 3.2.2.5$$

Equation 3.2.2.5 gives the Ideal Diode Equation by multiplying current density ‘J’ by cross-sectional area ‘A’ of the device:

$$I_D = I_{D0}\left[Exp\left(\frac{V_D}{V_{Th}}\right) - 1\right] \qquad 3.2.2.6$$

$$where\ I_{D0} = reverse\ leakage\ current\ or\ reverse\ saturation\ current$$
$$= A\left[q\frac{D_n}{L_n}n_p^- + q\frac{D_p}{L_p}p_n^-\right]$$

In Equation (3.2.2.5), Diffusion Coefficient and Diffusion Lengths are strongly dependent on the ambient temperature as well as the doping concentration of the P and N-Bulk.

Table 3.2.2.1. gives the typical values of Diffusion Lengths for different samples.

Table 3.2.2.1.Typical values of Diffusion Lengths for typical ssamples.

Sample	**Mobility of minority carrier(cm^2/V-s)**	**Life-Time**	**Diffusion Coefficient(cm^2/s)**	**Diffusion L**
P-TypeGaAs(300K)	4000	0.6ns	104(for electrons)	2.5µm
P-Type		0.01µs	25(for	5µm

Si(300K) N-Type Si(300K)	$N_D = 10^{17}$/cc	$\mu_n = 1100, \mu_p = 400$	1.44μs	electrons) 10.4(for holes)	38.69μm

3.2.2.1. Wide-Bulk Diode .

When the Bulk Widths are much greater than Diffusion Lengths of the minority carriers then the minority carriers perturbation have exponential decay and these Diodes are called Wide-Bulk Diode. The surface contacts or ohmic contacts donot influence the diode forward-bias current and donot effect the turn-on time.

Table 3.2.2.2. gives the typical parameters of an abrupt junction Wide-Bulk Diode.

Table 3.2.2.2.The fabrication parameters and the electrical parameters of a Wide-Bulk Abrupt Junction Diode.

N-Bulk	**P-Bulk**	Φ_{BO}	$\mathbf{d_n}$	$\mathbf{d_p}$	$\mathbf{d_0}$
10^{16}/cc	10^{18}/cc	0.834V	0.32μm	32A°	0.32μm

Table 3.2.2.3. gives the reverse saturation current or

reverse leakage current in a Wide-Bulk Diode.

Table 3.2.2.3.The carrier transport parameters and reverse saturation current of a Wide-Bulk diode with N-Type Doping 10^{16} /cc and P-Type Doping 10^{18} /cc and life-time $\tau_n = \tau_p = 1\mu s$, cross-sectional area $A = 10^{-3}\ cm^2$.

	N-side	**N-side**	**P-Side**	**P-Side**	**L_n on P-Side**	**L_p on N-Side**	**I_{DO}**
mobility	$\mu_p = 300(cm^2/V\ s)$	$\mu_n = 1200(cm^2/V\ s)$	$\mu_p = 100(cm^2/V\ s)$	$\mu_n = 280(cm^2/V\ s)$	0.27μm	0.279μm	0.01pA
Diff.coeff.	$D_p = 7.8(cm^2/s)$	$D_n = 31(cm^2/s)$	$D_p = 2.6(cm^2/s)$	$D_n = 7.3(cm^2/s)$			

For this Diode if we adopt Ideal Diode Equation then I-V characteristic is the following:

$$I_D = I_{D0}\left(Exp\left[\frac{V_D}{V_{Th}}\right] - 1\right) \quad where\ I_{D0} = 0.01pA \qquad 3.2.2.7$$

Plotting Equation 3.2.2.7. we get Figure 3.8 and Figure 3.9.

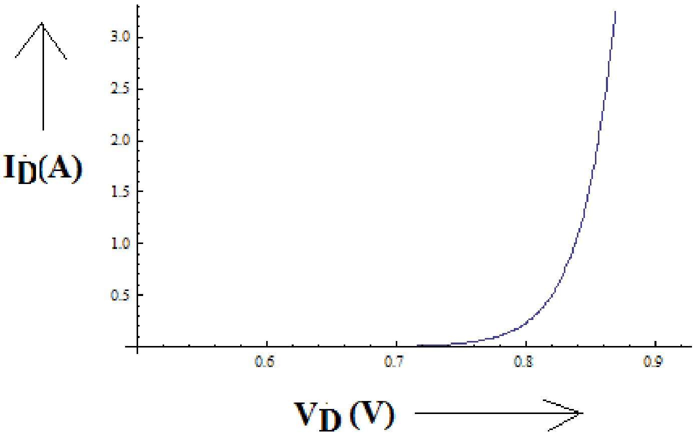


Figure 3.8. Forward I-V Characteristic of an Ideal Diode.

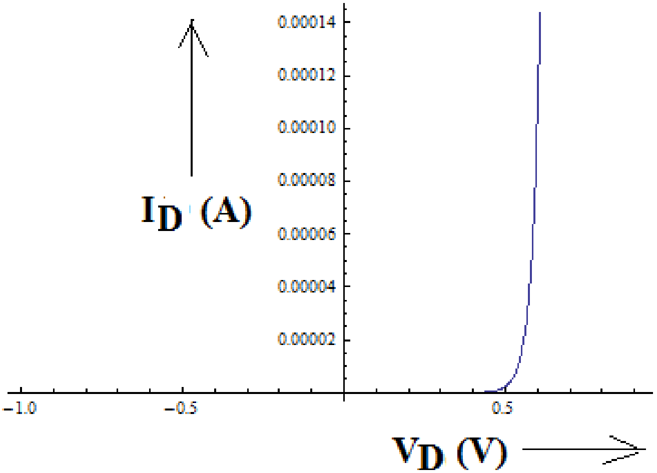


Figure 3.9. I-V characteristic of an Ideal Diode over the entire range of reverse(-1V) to forward bias(0.92V).

We clearly see that in Figure 3.9, the cut-in voltage of an Ideal Diode is 0.5V.

3.2.2.1. Narrow-Bulk Diode.

When N-Bulk and P-Bulk is much shorter than the corresponding minority carrier Diffusion Lengths then we have Narrow-Bulk Diode. In such a case we have linear decay of the minority carrier perturbation under forward bias condition as shown in Figure 3.10.

As can be seen in Figure 3.10, these new perturbations $p_n(0)$ at the edge of the depletion layer on N-side and $n_p(0)$ at the edge of the depletion layer on P-side decay linearly to thermal equilibrium values on their respective sides over the Bulk-width up to the ohmic contact where excess carrier is zero.

Under the condition shown n Figure 3.10. the ideal diode equation becomes:

$$J_D(forward) = \left[q\frac{D_n}{W_{BP}-d_p}n_p^- + q\frac{D_p}{W_{BN}-d_n}p_n^-\right]\left[Exp\left(\frac{V_F}{V_{Th}}\right)-1\right] \qquad 3.2.2.8$$

Equation 3.2.2.8 is identical to Equation 3.2.2.5 except for the length constant. The Diffusion Lengths of minority carriers have been replaced by effective Bulk-Widths on the respective sides.This results in the increase of reverse saturation current by a factor of 6.

If we take $W_{BP} = W_{BN} = 5\mu m$ which is much less than the diffusion lengths as can be seen in Table 3.2.2.3 then we get $I_{D0} = 5.6 \times 0.01pA$. But this gives a superior turn-on time as we will see in Diode Transients therefore narrow-bulk diode have applications.

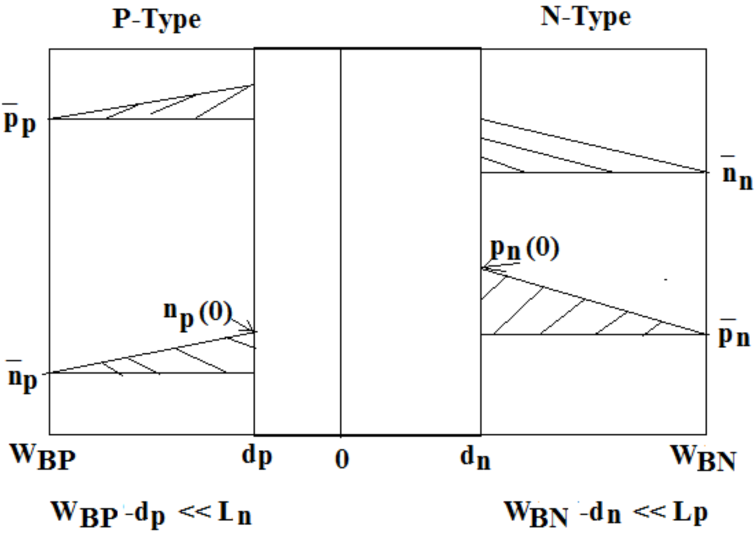


Figure 3.10. Minority and Majority triangular carrier concentration profile under forward bias in Narrow-Bulk Diode.

The theorization of narrow-bulk diode also helps understand the modern day BJT which are shallow junction and narrow bulk devices.

In the advanced version of Diode Physics we will discuss this narrow-bulk diode in a greater detail.

SSPD_Chapter 3_Section 3.2.3.Theoretical Formulation of the Knee Voltage(Vγ) of a forward biased-diode.

Section 3.2.3. gives the theoretical formulation of the knee voltage of a forward biased diode.

SSPD_Chapter 3_Section 3.2.3.Theoretical Formulation of the Knee Voltage(V γ) of a forward biased-diode.

In Section 3.2.2. we have derived the Ideal Diode Equation:

$$J_D(forward) = \left[q\frac{D_n}{L_n}n_p^- + q\frac{D_p}{L_p}p_n^-\right]\left[Exp\left(\frac{V_F}{V_{Th}}\right) - 1\right]$$

We choose the fabrication parameters of a typical Diode as given in Table 3.2.2.3. They are as follows:

Wide-Bulk diode with N-Type Doping 10^{16}/cc and P-Type Doping 10^{18}/cc and life-time $\tau_n = \tau_p = 1\mu s$, cross-sectional area $A = 10^{-3} cm^2$.

D_n(on P-side) = $7.3 cm^2/s$; L_n(on P-side) = $0.27\mu m$;

D_p(on N-side) = $7.8 cm^2/s$; L_p(on N-side) = $0.279\mu m$;

Therefore $I_D = I_{D0}[\lambda - 1]$ where $\lambda = Exp(V_F/V_{Th})$;

Substituting the fabrication parameters we get:

$I_D = I_{D0}[\lambda-1] = 4.5\times 10^{-13}[\lambda-1]$Amp.

Let us define the knee voltage of a forward biased diode as the forward bias at which 10μA flows through the diode. Hence Ideal Diode Equation is:

$$I_D(forward) = 10\mu A = A(cm^2)\left[q\frac{D_n}{L_n}n_p^- + q\frac{D_p}{L_p}p_n^-\right]\left[Exp\left(\frac{V_\gamma}{V_{Th}}\right) - 1\right] \qquad 3.2.3.1$$

Since Thermal Voltage is 26mV and Knee Voltage will be several times greater than V_{Th} therefore Equation 3.2.3.1 reduces to:

$$10\mu A = A(cm^2)\left[q\frac{D_n}{L_n}n_p^- + q\frac{D_p}{L_p}p_n^-\right]\left[Exp\left(\frac{V_\gamma}{V_{Th}}\right)\right] \qquad 3.2.3.2$$

Thermal equilibrium value of minority carrier can be replaced by (n_i^2/dopent density). Equation 3.2.3.2 becomes:

$$10\mu A = A(cm^2)\left[q\frac{D_n}{L_n}\times\frac{1}{N_A} + q\frac{D_p}{L_p}\times\frac{1}{N_D}\right]n_i^2\left[Exp\left(\frac{V_\gamma}{V_{Th}}\right)\right] \qquad 3.2.3.3$$

Replacing n_i^2 by its theoretical form given in Equation(3.2.11) we get:

$$10\mu A = A(cm^2)\left[q\frac{D_n}{L_n}\times\frac{1}{N_A} + q\frac{D_p}{L_p}\times\frac{1}{N_D}\right]N_C N_V\left[Exp\left(-\frac{V_g}{V_{Th}}\right)\right]\left[Exp\left(\frac{V_\gamma}{V_{Th}}\right)\right] \qquad 3.2.3.3$$

Substituting the fabrication and electrical parameters we get:

$$10\mu A = 45225\left[Exp\left(-\frac{V_g}{V_{Th}}\right)\right]\left[Exp\left(\frac{V_\gamma}{V_{Th}}\right)\right] \qquad 3.2.3.3$$

Equation (3.2.3.3) simplifies to :

$$\frac{10\mu A}{452252} = Exp\left[-\frac{(V_g - V_\gamma)}{V_{Th}}\right] \qquad 3.2.3.4.$$

Taking the natural log of Equation (3.2.3.4):

$$-24.3 = \frac{-V_g + V_\gamma}{V_{Th}} \qquad 3.2.3.5.$$

Therefore $V_\gamma = V_g - 24.53 \times V_{Th} = V_g - 0.6379$ ______________________ 3.2.3.6.

The Knee Voltages of Ge, Si and GaAs at 300K are tabulated in Table 3.2.3.1.

Table 3.2.3.1.Knee Voltages of Ge,Si and GaAs diodes at 300K frpm Equation (3.2.3.6)

Material	**Band-Gap Voltage(V)**	**Knee Voltage(V)**
Ge	0.67	0.03
Si	1.12	0.482
GaAs	1.4	0.7621

By observation the knee voltages observed for the three diodes are of the same range as given in the Table. From Equation (3.2.3.6) we see that knee voltage is determined by material Enegy Band Gap Voltage.

SSPD_Chapter 3_SEc3.2.4. Reverse biasedDiode,3.2.5.Real Diodes,3.2.6.Avalanche Diode,3.2.7.Hyperabrupt Jn.Diode.
Section3.2.4, 3.2.5,3.2.6 and 3.2.7 compares the real diode with ideal diode, its temperature dependence, Avaalanche Break down under reverse bias condition and design of Varactor Diode(hyperabrupt diode).

3.2.4.Reverse Biased Diode.

In Section 3.2.2. we have derived the Shockley Equation or the Ideal Diode Equation from the first principles of the forward biased diode. In the present section the same Ideal Diode Equation will be verified by studying the reverse biased diode. This analysis will give the physical origin of the reverse leakage current or the reverse saturation current. In Figure 3.11. the space charge region, the electric field and the potential gradient in a reverse biased diode is given.

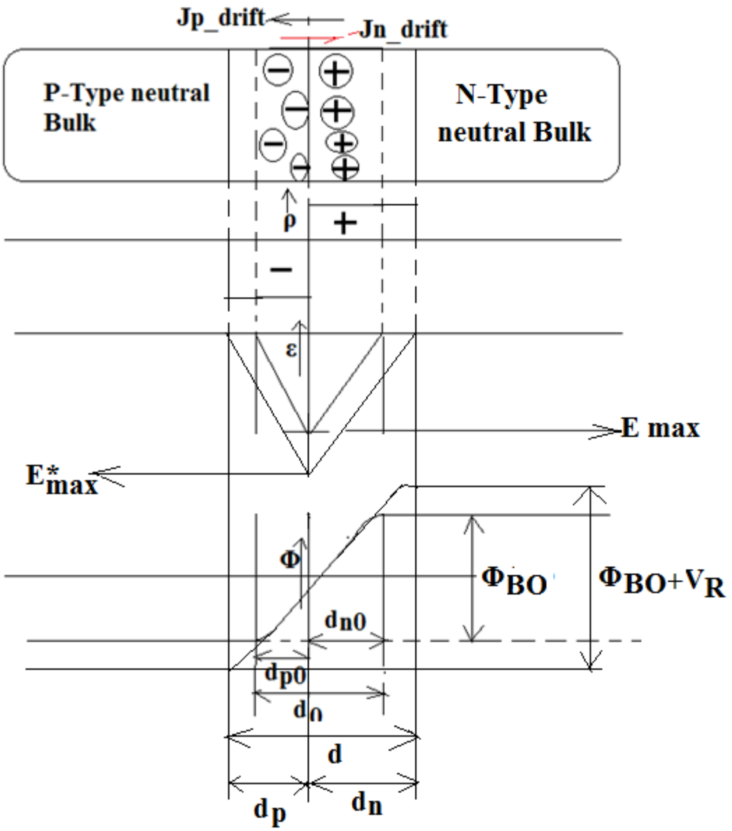


Figure 3.11. Space Charge Profile, Electric Field Profile and the Potential Profile along the longitudinal axis of a reverse biased diode.

Under reverse bias, as seen from Figure 3.11., the depletion region widens, maximum Electric Field at the junction becomes more intense and built-in barrier potential is now ($\Phi_{BO} + V_R$).

Equation 3.2.9. gets modified to the following form:

$$d = \sqrt{\frac{2 \in}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right)} \times \sqrt{\emptyset_{BO} + V_R} \qquad 3.2.4.1$$

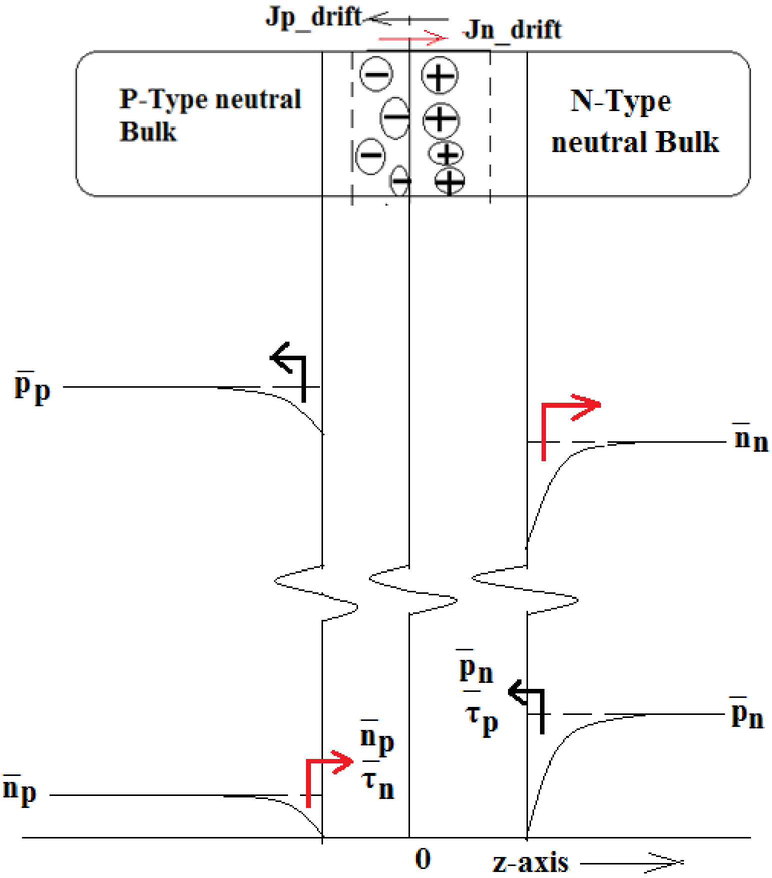


Figure 3.12. Carrier Concentration profile of the majority and minority carriers along the Z-axis.

Diffusion current of majority carriers is completely stopped because of large barrier potential. There is only drift of the minority carriers in the reverse direction.By examining Figure 3.12 we can analytically deduce the drift currents and hence the total reverse leakage current. In Figure 3.12. the carrier concentration of a reverse biased diode is shown.

By Boltzmann relation,at the edge of the depletion region, minority carrier concentration is perturbed in the following manner:

$$p_n^- \rightarrow p_n^- Exp\left(\frac{V_D}{V_{Th}}\right); n_p^- \rightarrow n_p^- Exp\left(\frac{V_D}{V_{Th}}\right); \qquad 3.2.4.2$$

Equation (3.2.4.2) tells us that under forward bias there is excess of carriers at the edge of the depletion region hence there is recombination and under reverse bias there is deficit of carriers and hence there is thermal generation of EHP and the the minority carrier is suppressed to zero at the edge of the depletion region hence the generation rate is:

$$Generation\ of\ minority\ carriers\ on\ P-Side = \frac{n_p^-}{\tau_n}\ per\ second\ per\ cc.$$

$$Generation\ of\ minority\ carriers\ on\ N-Side = \frac{p_n^-}{\tau_p}\ per\ second\ per\ cc.$$

All the generated carriers within L_n (diffusion length of electrons) manage to survive and reach the edge of the depletion region on the P-Side. Similarly All the generated carriers within L_p (diffusion length of holes) manage to survive and reach the edge of the depletion region on the N-Side.Hence total reverse saturation current or reverse leakage current is as follows:

$$I_{DO} = \frac{n_p^-}{\tau_n} \times Aq \times L_n + \frac{p_n^-}{\tau_p} \times Aq \times L_p \qquad 3.2.4.3$$

This is the same as the reverse leakage current in Equation (3.2.2.6).

The reverse leakage current in Equation (3.2.2.6) is:

$$I_{DO} = A\left[q\frac{D_n}{L_n}n_p^- + q\frac{D_p}{L_p}p_n^-\right] \qquad 3.2.4.4$$

$$Since\ L_n^2 = D_n\tau_n\ and\ L_p^2 = D_p\tau_p\ \ therefore\ \frac{L_n}{\tau_n} = \frac{D_n}{L_n}\ and\ \frac{L_p}{\tau_p} = \frac{D_p}{L_p};$$

Hence the Equations (3.2.4.3) and (3.2.4.4) are equivalent and they are due to the limited minority carriers being generated at the two sides of the depletion region. As soon as they are generated they drift down the potential gradient of the built-in barrier potential.

Since they depend upon the generation of the minority carriers hence they are very temperature sensitive and they double for every 10°C rise in temperature for Si as well as for Ge diodes.

3.2.4.1.Temperature dependence of I DO (reverse leakage current).

Rewriting Equation(3.2.4.4):

$$I_{DO} = A\left[q\frac{D_n}{L_n}n_p^- + q\frac{D_p}{L_p}p_n^-\right] \qquad 3.2.4.4$$

Therefore:

$$I_{DO} = Aq\left[\frac{D_n}{L_n}\times\frac{n_i^2}{N_A} + \frac{D_p}{L_p}\times\frac{n_i^2}{N_D}p_n^-\right] = Aq\left[\frac{D_n}{L_n}\times\frac{1}{N_A} + \frac{D_p}{L_p}\times\frac{1}{N_D}p_n^-\right]n_i^2 \qquad 3.2.4.5.$$

From Equation (3.2.4.5), it is evident that for small increments if we assume mobility and diffusion coefficient to be independent of temperature then I_{DO} varies directly as n_i^2.

$$n_i^2 = N_C N_V Exp\left[-\frac{E_g}{kT}\right] therefore \frac{n_i^2(T+1)}{n_i^2(T)} = Exp\left[\frac{E_g}{kT(T+1)}\right] = 1.1548 \quad 3.2.4.6$$

Equation tells us that ni2 increases by 15.48% for every 1K rise in Temperature.

But experimentally, ni2 increases by 7% for every 1K rise in Temperature for both Si and Ge Diodes.

Hence we can assume that IDO increases by 7% for every 1°C rise in Temperature.

Therefore:

$$I_{DO}(T+1) = 1.07 I_{DO}(T) \qquad 3.2.4.5.$$

If for 1°C rise in temperature, IDO increases by 1.07 then for 10°C rise in temperature, IDO increases by (1.07)10 = 1.969 ~ 2.

Therefore we can say that Reverse Saturation Current doubles for every 10°C rise in Ambient Temperature.

Hence the expression for Reverse Saturation Current is:

$$I_{DO}(T_x) = 2^{\frac{T_x - T_{amb}}{10}} I_{DO}(T_{amb}) \qquad 3.2.4.6.$$

3.2.4.2.Temperature Dependence of Forward Bias Voltage V F for maintaining a constant forward current I F .

Since ID0 doubles for every 10°C rise in temperature hence forward current also increases as shown in Figure 3.13. Therefore at Room Temperature, forward bias must decrease by 2.5mV for every 1°C rise in temperature to maintain forward current constant.

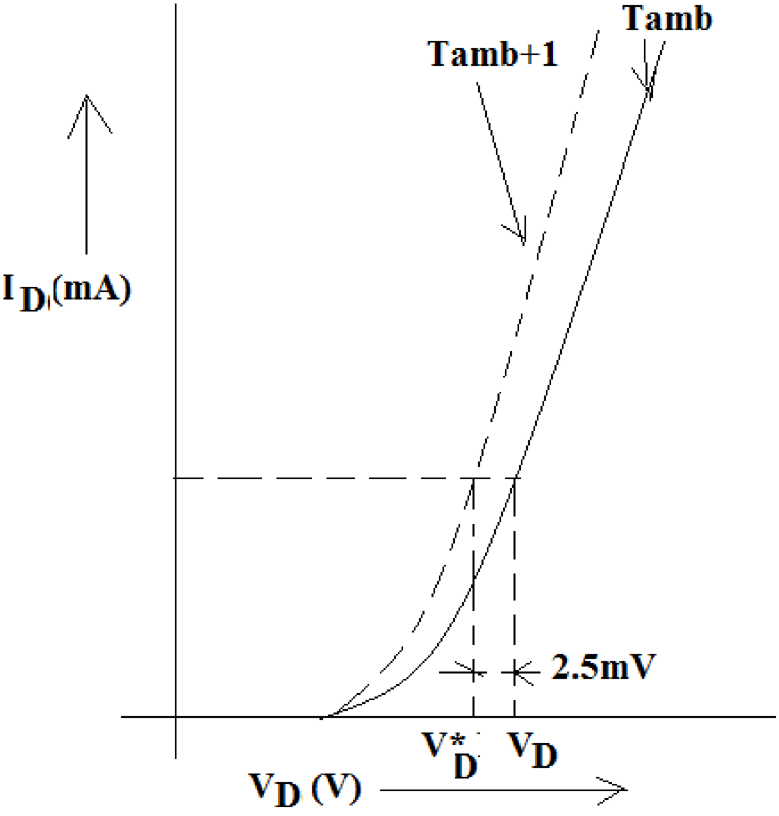


Figure 3.13.Forward Voltage reduction by 2.5mV for every 1 degree centigrade rise in temperature.

At Temperature T Kelvin:

$$I_D(T) = I_{D0}(T) Exp\left[\frac{V_D}{kT}\right] \quad 3.2.4.2.1.$$

At Temperature (T + 1) Kelvin:

$$I_D(T+1) = I_{D0}(T+1)Exp\left[\frac{V_D}{k(T+1)}\right] \tag{3.2.4.2.2.}$$

But $I_{D0}(T+1) = 1.07 I_{D0}(T)$ hence if the current has to remain constant then V_D must reduce by ΔV.

Therefore Equation (3.2.4.2.2) is rewritten for constancy of the current:

$$I_{D0}(T)Exp\left[\frac{V_D}{kT}\right] = 1.07 I_{D0}(T)Exp\left[\frac{V_D - \Delta V}{k(T+1)}\right] \tag{3.2.4.2.3}$$

Simplifying Equation (3.2.4.2.3) we get:

$$Exp\left[\frac{V_D}{kT} - \frac{V_D - \Delta V}{kT(1+\frac{1}{T})}\right] = 1.07 \tag{3.2.4.2.4}$$

At T = 300K, 1/300 < < 1 therefore Equation (3.2.4.2.4) simplifies to:

$$\frac{\Delta V}{kT} = \ln(1.07)$$

Or

$$\Delta V = kT\,ln(1.07) = 26mV \times 0.0677 = 1.759mV \tag{3.2.4.2.5}$$

Theoretically forward voltage must decrease by 1.759mV/1°C to maintain I_D constant but exoerimentally:

$$\frac{dV_D}{dT}\Big|_{I_D} = -\frac{2.5mV}{K}\ for\ both\ Si\ and\ Ge\ Diodes \tag{3.2.4.2.6.}$$

3.2.5.Real Diode Equation.

Two simplifying assumptions were made while deriving Shockley Equation:

i.Bulk is quasi-neutral. This means the applied diode voltage drops across the depletion region only.

ii.There is no recombination in the depletion region.

In real diodes, at low voltages the depletion region is comparable to diffusion lengths hence while crossing the depletion region recombination does take place. As a result the current is not constant while crossing the depletion region.

Therefore an Ideality Factor has to be introduced in the Ideal Diode Equation. That is below 0.5mA, Ideal Diode Equation is not adequate to describe the real diode behavior.

Theefore at low currents:

$$I_D = I_{D0}\left[Exp\left(\frac{V_D}{\eta V_{Th}}\right) - 1\right] \qquad 3.2.5.1$$

Where Ideality Factor η =2 at low currents lower than 0.5mA.

At high currents, when we have high injection level then the bulk does not remain neutral. At high injection levels, perturbation in minority carrier becomes comparable to the majority carrier concentration that is:

$$p_n^{\wedge} \sim n_n^{-} \quad and \; n_p^{\wedge} \sim p_p^{-}$$

$$p_n^{\wedge} = p_n^{-} Exp \frac{V_D^*}{V_{Th}} \sim N_D$$

Therefore the forward voltage where High Level Injection(HIL) occurs:

$$V_D^* \sim V_{Th} ln \frac{N_D}{p_n^-} = V_{Th} ln \frac{N_D^2}{n_i^2} = 2V_{Th} ln \frac{N_D}{n_i} \sim 0.83V \; for \; N_D = \frac{10^{16}}{cc}; \qquad 3.2.5.2$$

Under such a condition the carrier concentration profile as shown in Figure 3.14:

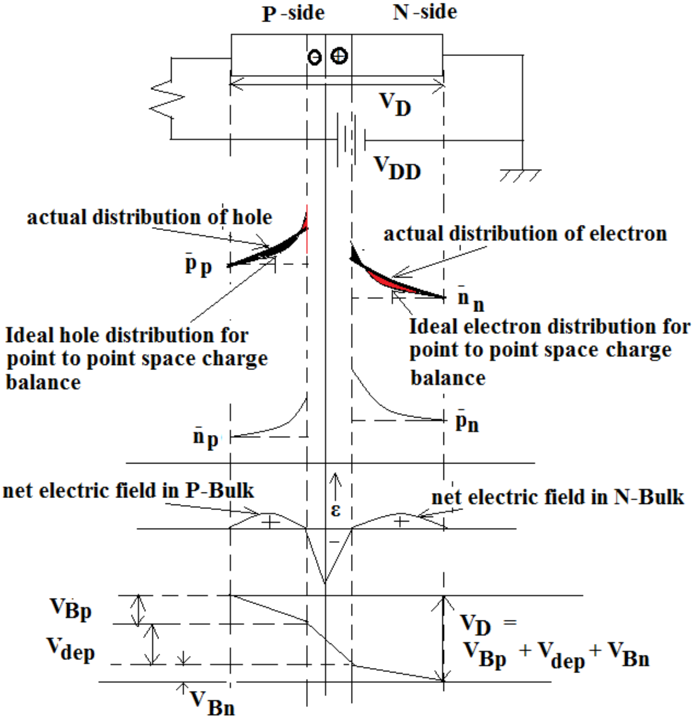

P-side
N-side
V_D
V_{DD}
actual distribution of hole
$\bar{p}_p$
actual distribution of electron
$\bar{n}_n$
Ideal hole distribution for point to point space charge balance
Ideal electron distribution for point to point space charge balance
$\bar{n}_p$
$\bar{p}_n$
net electric field in P-Bulk
ε
net electric field in N-Bulk
+
−
+
V_{Bp}
V_{dep}
$V_D = V_{Bp} + V_{dep} + V_{Bn}$
V_{Bn}

Figure 3.14.The Votage Drop across the bulk and the depletion region under HIL.

As seen in Figure 3.14. there is considerable field and hence voltage drop across the Bulk. Therefore

$V_D = V_{dep}$ is invalid now.

$V_{dep} = V_D - (V_{Bp} + V_{Bn})$;

It will be shown through detailed analysis in the Appendix of Chapter 3 that

$V_{dep} = (1/2)V_D$. Since the drop across the depletion region decides the Shockley Equation hence at HIL the form of diode equation is:

$$I_D = I_{D0}\left[Exp\left(\frac{V_D}{\eta V_{Th}}\right) - 1\right]$$

Where η = 2.

This discussion tells us that Real Diodes are described by Equation 3.2.5.1

where η = 2 for currents less than 0.5mA(depletion region recombination dominates)

and η = 2 for currents more than 5mA(the bulk does not remain quasi-neutral).

If we assume that diode is on then ‘1’ can be neglected and equation is expressed as:

$$I_D = I_{DO}Exp\left[\frac{V_D}{\eta V_{Th}}\right] = I_{D0}10^{\frac{V_D}{\eta 60mV}} \qquad 3.2.5.3$$

This Equation (3.2.5.3) tells us that for every 60mV increase in Diode Voltage increases the Diode

Current by 10 times that is by a decade under moderate current where η = 1.

For low currents and for high currents where η = 2, current increase by a decade for 129mV increase in Diode Voltage.

For Ideal Diode the semi-log plot of Ideal Diode Equation gives 60mV/decade but for Real Diode we get 120mV per decade slope at currents less than 0.5mA, 60mV/decade slope for intermediate currents 0.5mA to 5mA and again 120mV/decade slope for currents in the range of 5mA and above. Semi-log plot of a Real-Diode is shown in Figure 3.15.

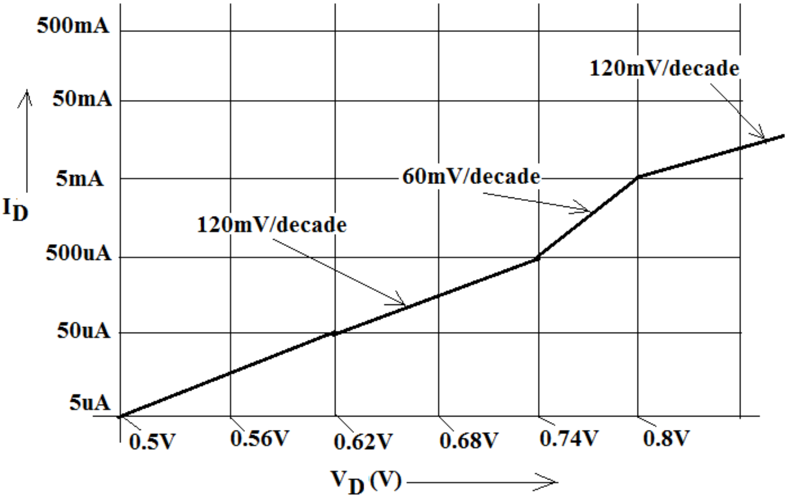


Figure 3.15. Semi-Log Plot of I-V characteristic of a Real-Life Diode.

3.2.6. Avalanche Breakdown .

Under reverse bias condition when E_{max} exceeds

E_{crit} (constant at 3×10^{5}V/cm) also known as Breakdown field of Silicon Diode under low doping condition then Avalanche Breakdown takes place and the diode can be irreversibly damaged if there is no limiting resistance in the circuit.

As Reverse Bias Voltage starts increasing the reverse leakage current starts multiplying by an Avalanche Mutiplication Factor ‘M’ where M is given as follows:

$$M = \frac{1}{1-(\frac{V_R}{BV_A})^n} \qquad where\ n = Miller\ indices = 1\ to\ 3 \tag{3.2.6.1.}$$

The reverse leakage current is given by the following expression:

$$I_R = -MI_{D0} \tag{3.2.6.2.}$$

As the reverse voltage approaches BV_A(Avalanche Break-Down Voltage), the drifting minority carriers get accelerated to a point where they break the co-valent bond and EHPs are produced by impact collision and ionization.Soon this becomes a runaway process and current rapidly builds up only to be limitesd by the circuit loop resistance.

For one sided step junction (P^+N Junxction Diode), depletion layer extends into the lightly doped side and electric field is given by:

$$E_{max} = \frac{qN_D d_n}{\epsilon} = E_{crit} \tag{3.2.6.3}$$

Therefore Avalanche Breakdown Voltage:

$$BV_A = \frac{E_{crit} d_n}{2} = \frac{\epsilon E_{crit}^2}{2q N_D} = 28V \ for \ N_D = \frac{10^{16}}{cc} \qquad 3.2.6.4$$

The corresponding depletion layer is: d_n = 1.9μm.

This picture holds true only for lightly doped diodes. At higher doping Quantum Mechanical Tunneling comes into picture.

3.2.7.Hyperabrupt Junction Diodes to synthesize Varactor Diode.

We saw in the fabrication section that doping profile can be complimentary error function from a short period constant source diffusion and Gaussian Distribution for a long period limited source – diffusion.In general the doping profile is the following:

$$doping = K z^m \qquad 3.2.7.1$$

For this generalized doping we get a generalized power function dependence of Junction Capacitance. The junction Capacitance of a diode is given by the same formula as the parallel plate Capacitor:

$$C_{junction} = \frac{\epsilon A}{d} \ where \ A = cross - sectional \ area \ and \ d = depletion \ width.$$

By detailed analysis it has been shown that :

$$C_{junction} = K_1 V_R^{-n} \ where \ n = \frac{1}{m+2} \qquad 3.2.7.2$$

Using Equation (3.2.7.1) and (3.2.7.2), abrupt junction diode, linealrly graded junction diode and hyper-abrupt-junction diode can be designed. Table 3.2.7.1 gives the Junction Capacitance Power Law.

Table 3.2.7.1. Junction Capacitance Power Law for generalized doping.

Type of junction	m	n	Power Law	Application
Abrupt Jn.	0	1/2	Inverse square root law	General Application
Linearly graded Jn.	1	1/3	Inverse cube root law	For special application
Hyper-abrupt Jn.	-3/2	2	inverse as the square of Voltage	Voltage Controlled Oscillator

The doping profile of the three kinds of the junctions are given in Figure 3.16.

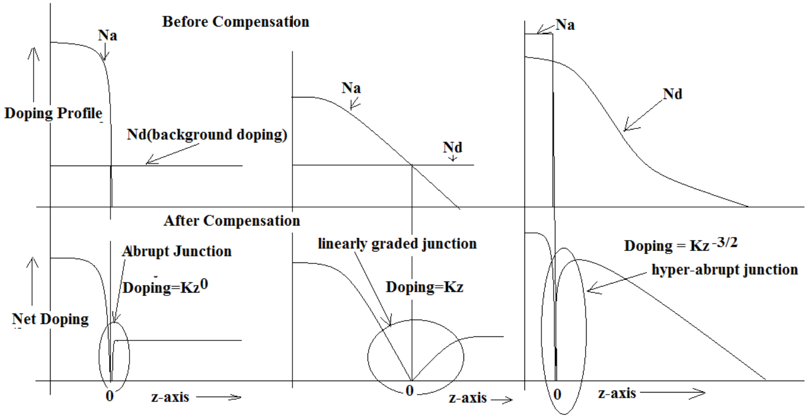


Figure 3.16. Impurity Doping Profile and the resultant junction.

In the next section we will examine hyper-abrupt Junction and its application in Voltage Controlled Oscillator as a VARACTOR DIODE which is commonly used in FM generation..

SSPD_Chapter 3.2.7.1.. Application of Hyperabrupt Junction Diode as Varactor Diode in Voltage-controlled Oscillator(VCO).
Section 3.2.7.1. gives the application of hyper-abrupt junction diode as Varactor Diode used in Voltage Controlled Oscillator.

3.2.7.1. Application of Hyperabrupt Junction Diode as Varactor Diode in Voltage-controlled Oscillator(VCO).

In Equation (3.2.7.2) we saw that in hyperabrupt Junction diode, junction capacitance (C_j) varies inversely as the square of the Reverse Junction Bias Voltage (V_j).

Hence:

$$C_j = \frac{\epsilon A}{K V_j^2} \; where\ A = cross - sectional\ area\ and\ K = constant\ of\ equat. \quad 3.2.7.3$$

If the diode with this unique power law is utilized in a parallel resonance circuit then the resonance frequency ω_0 becomes directly proportional to the diode voltage as shown in subsequent paragraphs.In this set up hyper-abrupt junction diode acts a variable reactance device hence it is called Varactor Diode and application of Varactor Diode enables the synthesis of Voltage Controlled Oscillator(VCO) which is one of the most popular method of Frequency Modulation of a Carrier Signal in FM

Radio Broadcast System.

In Figure 3.17, VCO using a Varactor Diode is illustrated.

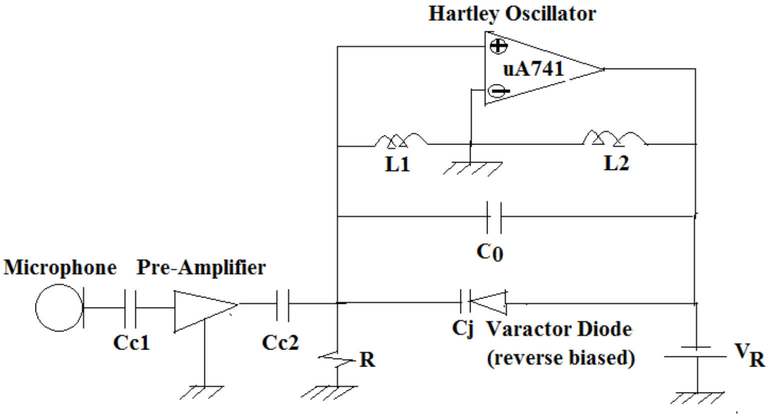


Figure 3.17. Application of Varactor Diode in synthesizing VCO.

Theoretical Formulation of VCO.

In the Hartley Oscillator shown in Figure 3.17., the frequency of oscillation is as follows:

$$\omega_{carrier} = \frac{1}{\sqrt{L_0(C_0+C_j)}} = \frac{1}{\sqrt{L_0C_0\left(1+\frac{C_j}{C_0}\right)}} = \frac{1}{\sqrt{L_0C_0}} \times \sqrt{\frac{C_0}{C_j}} \quad assumimng\ \frac{C_j}{C_0} \gg 1;$$

Therefore:

$$\omega_{carrier} = \omega_0\sqrt{\frac{C_0}{C_j}} = \frac{\omega_0\sqrt{C_0}}{\sqrt{\frac{\epsilon A}{K}}}(V_j) \quad where\ \omega_0 = \frac{1}{\sqrt{L_0C_0}} \qquad 3.2.7.4$$

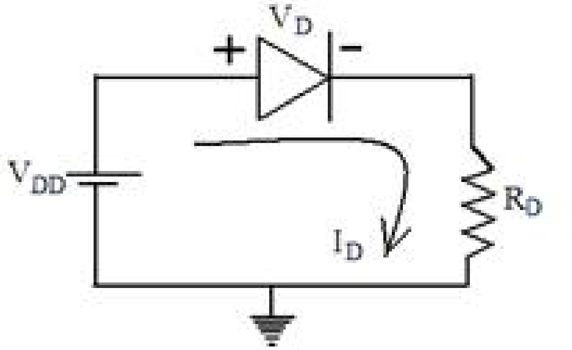
Equation (3.2.7.4) gives the theoretical basis of VCO. The voltage generated from the microphone is pre-amplified and then coupled to the reverse bias loop of the varactor diode. The varactor diode junction capacitance forms a part of the parallel resonance circuit and the audio voltage generated from the microphone directly controls the carrier frequency generated by the Hartley Oscilator. Thus we have achieved a VCO where audio signal of the Radio Broadcast is carrying out the Frequency Modulation of the carrier frequency generated by the Hartley Oscillator.

AnalogElectronics_Lecture3_Incremental Model of Diode.
This develops the low frequency incremental model of diode and illustrates the harmonic distortion when small signal condition is violated.

LECTURE NO. 3

INCREMENTAL MODEL OF A PN JUNCTION DIODE

Fig.1. A dc diode circuit.



$V_{DD} = V_D + I_D . R_D$___(1) This describes DC load line.

$I_D = I_{D0} \exp(V_D/V_T)$___(2) This is the device characteristics.

Q-point or the quiescent point is the DC operating point and is obtained as the intersection of DC load line and the device characteristics.

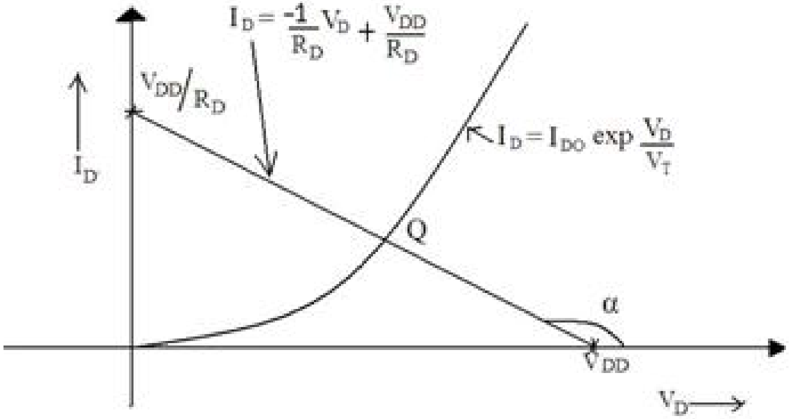


Figure 2. The dc load line , the device characteristics and the Q point.

Tan(α)= slope of the load line= (-1/RD)

Under signal conditions:

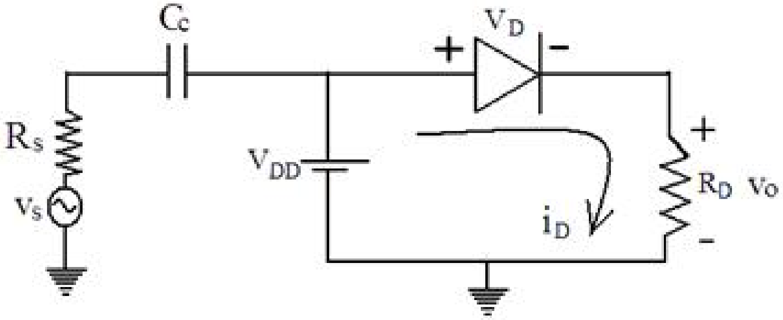


Figure 3. Signal is being coupled with the diode circuit.

$i_D = I_D + i_d$ where i_D is the instantaneous diode current.

I_D is the DC diode current.

i_d is the incremental diode current.

And v_D (instantaneous diode voltage) = V_D (DC Diode Voltage) + v_d_(incremental diode voltage)

Now the loop or the mesh equation is:-

$(V_{DD} + v_s) = v_D + i_D R_D$_____(3)

Rewriting the above equation we get:-

$(V_{DD} + v_s) = (V_D + v_d) + (I_D + i_d) R_D$____(4)

Now we have (Instantaneous – DC) = Incremental part,

that is Eq.(4)-Eq.(1) :

$v_s = v_d + i_d R_D = v_d + v_o$;

Incremental circuit will be:-

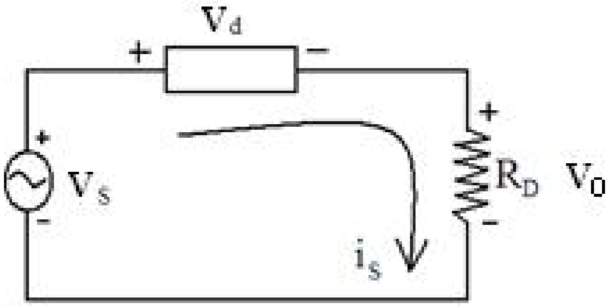
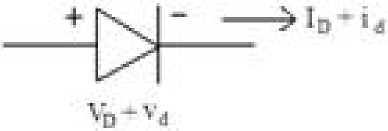


Figure 4. The incremental circuit of the Diode Circuit with signal.

A diode under instantaneous conditions has two parts :-



$$(I_D+i_d) = I_{DO}\exp\left(\frac{V_D+v_d}{V_T}\right)$$

The above is a relation between the diode current and the diode voltage.

We know that

$e^{\Theta} = 1 + \Theta + (\Theta^2/2!) + (\Theta^3/3!) + \ldots\ldots\ldots\ldots$

If $\Theta << 1$; $e^{\Theta} = 1 + \Theta$ This now becomes a linear equation.

If incremental voltage across the diode is less than 5 mV then $\Theta << 1$.

To maintain linearity, we maintain all the signals small.

So under small signal approximations,

$e^{\Theta} = 1 + \Theta$ will hold good.

$$(I_D + i_d) = I_{DO}\left[\exp\left(\frac{V_D}{V_T}\right)\right]\left[\exp\left(\frac{V_d}{V_T}\right)\right]$$

Now we note that $V_T = 26\text{mV}$ and the room temperature $= 300\text{K}$ and if $V_d < 5$ mV then we satisfy the small signal condition and we obtain:

$$(I_D + i_d) = I_D\left[1 + \frac{V_d}{V_T}\right]$$

Then we have;

$$\cancel{I_D} + i_d = \cancel{I_D} + I_D \frac{v_d}{V_T}$$

Thus;

$$i_d \frac{V_T}{I_D} = v_d$$

Let:

$$\frac{V_T}{I_D} = r_d$$

Putting $V_T = 26\text{mV}$ and $I_D = 1\text{mA}$ we get $r_d = 26\Omega$.

Thus the incremental part of the diode circuit was determined as follows:

- We short circuit the DC voltage source(DC current source would have been open circuited).
- Any forward biased junction would be replaced by incremental resistance.

Incremental resistance r_d = (Thermal Resistance

V_T/ Quiescent Current through the diode)

$$r_d = \frac{V_T}{I_D}$$

$$r_d = \frac{1}{\text{Slope of the tangent drawn at the Q-point}}$$

- All backward bias junction will act as open circuit.
- Ohmic resistance will offer resistance in both DC and incremental part.

This is small signal approximation. Thus Diode equivalent circuit is composed of linear elements only under small signal condition. Hence the circuit in Figure 4 is incremental circuit or small signal equivalent circuit. The incremental resistance r_d offered by the diode under small signal condition is a linear resistance and is included in the circuit only under incremental condition.

Table 1. Values of incremental resistance at 300K offered by a diode under various DC diode currents.

I_D (mA)	r_d (ohms) = V_T/I_D at Room Temerature(300K
1µA	26 kohms
10µA	2.6 kohms
100µA	260 ohms
1 mA	26 ohms
10mA	2.6ohms
100mA	0.26 ohms.

Amplitude or Harmonic Distortion.

As can be seen from Figure 3, input sinusoidal voltage v_S appears as addition and subtraction to V_{DD}. Hence under signal condition, the load line is being shifted as shown in Figure 5. In doing so Q pint also shifts generating sinusoidal current swing. As can be seen in the figure, a small segment of I-V curve of the diode(which is the case under small signal condition) is essentially a straight line. Hence Q moves along a straight segment and in the process generates a sinusoidal swing in the current.

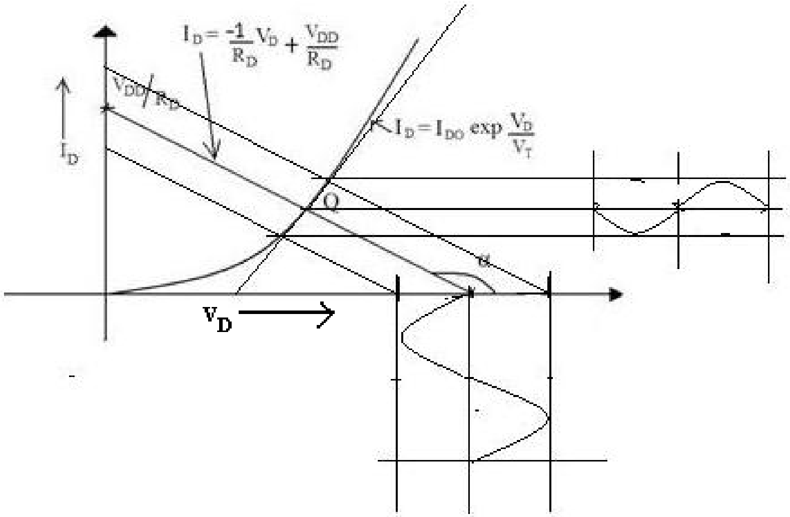

$I_D = \frac{-1}{R_D} V_D + \frac{V_{DD}}{R_D}$
V_{DD}/R_D
I_D
$I_D = I_{D0} \exp \frac{V_D}{V_T}$
Q
α
V_D

Figure 5. Diode Current Sinusoidal Swing under Sinusoidal Input Voltage under small signal condition.

But as seen in Figure 6, if input voltage is a large signal then Q traces a non-linear segment of I-V curve. This results in non-sinusoidal current swing in the diode. This means a sinusoidal voltage is not giving rise to a sinusoidal current in the output load resistance RL. Hence the output voltage will be non-sinusoidal and its Fourier Series Expansion will contain Fundamental and Harmonics. This is known as Amplitude or Harmonic Distortion.

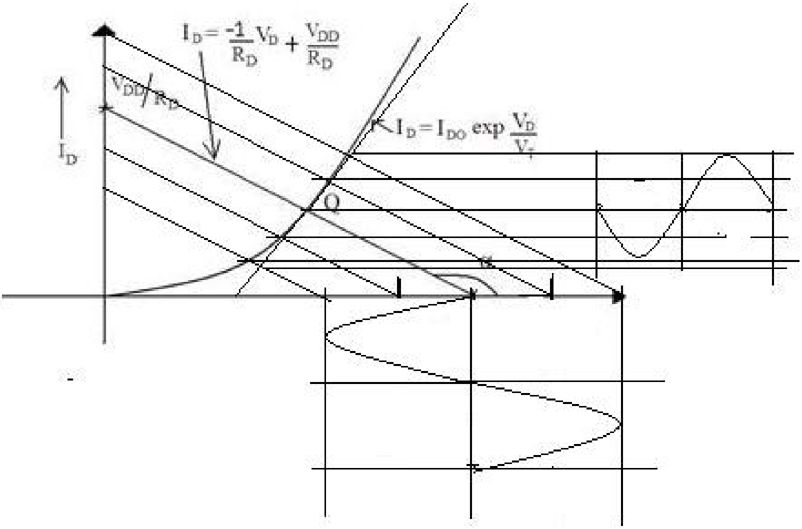

$I_D = \frac{-1}{R_D} V_D + \frac{V_{DD}}{R_D}$
V_{DD}/R_D
I_D
$I_D = I_{DO} \exp \frac{V_D}{V_T}$
Q

Figure 6. Non-sinusoidal diode current swing when input voltage is a large voltage.

AnalogElectronics_Lecture3supp._High Frequency Model of PN Junction Diode.

This describes the high frequency model of diode. It gives the physics and theoretical formulation of the junction capacitance and diffusion capacitance.

AnalogElectronics_Lecture3supp._High Frequency Model of PN Junction Diode.

At high frequency, the Depletion Layer Junction Capacitance(C_{JD}) under reverse biased and {C_{JD} plus C_D(Diffusion Capacitance)} comes in parallel with the incremental resistance r_d . At very high frequencies even the leads will offer an inductive reactance but we are going to neglect it for the present syllabus. The incremental Model of PN Junction Diode at high frequencies is given in Figure 1.

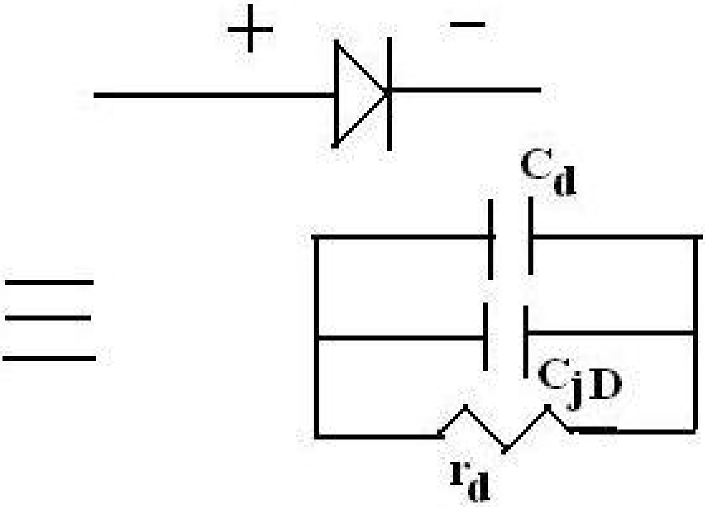


Figure 1. Incremental Model of Diode at high frequencies.

Junction Capacitance (C_{jD}) is present in both reverse and forward biased diodes because it is due to depletion layer/space charge layer/dipole layer present at the metallurgical junction of the diode.

Diffusion Capacitance(C_d) is present only in forward bias because it is due to excess minority carrier stored under forward biased condition. In reverse bias condition there is no excess minority carrier stored in the bulk region.

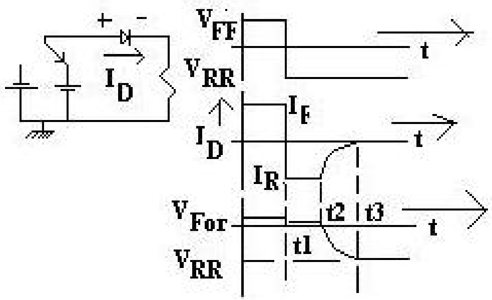
In Figure 2 we see the deleterious effect of $C_d + C_{jD}$ on the switching performance of the diode. Ideally a diode should stop conducting and it should switch off but this switch off is not instantaneous. If a forward biased diode is reverse biased it continues to conduct in the reverse direction for time ($t_{storage} + t_{discharge}$) before it switches off. The switching transient of a diode is shown in Figure 2.

The delay in switching is due to the time delay in removing the excess minority carriers and time taken in discharging the junction capacitance.

Time delay in removing the minority carriers is storage delay = $t_{storage}$;

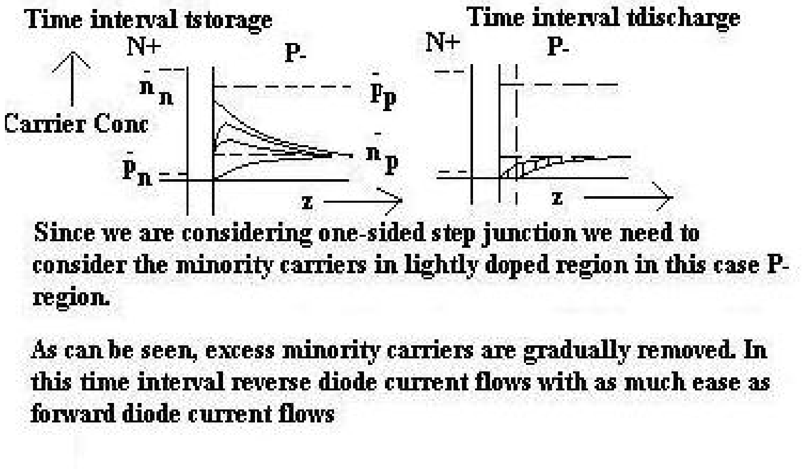
Time taken in discharging the junction capacitance shows up as a fall time = $t_{discharge}$;

The Physics of removal of minority carriers and discharging of C_{JD} are shown in Figure 3.



A diode circuit which is initially forward biased. At time t1, by a Single Pole Double Throw Switch(SPDT), it is abruptly reverse biased. Note that diode does not abruptly switch off.

Figure 2.Switching Transient of a Diode.



Since we are considering one-sided step junction we need to consider the minority carriers in lightly doped region in this case P- region.

As can be seen, excess minority carriers are gradually removed. In this time interval reverse diode current flows with as much ease as forward diode current flows

Figure 3. The process of minority carrier removal during (t storage + t discharge)

Diode Junction Capacitance (CjD) = ($\varepsilon_0\varepsilon_r A/d$) where A= cross sectional area of the diode and d=depletion width;

Physics of Diode Junction Capacitance:

The doping profile of N+ and P- decide the voltage dependence of Junction Capacitance.

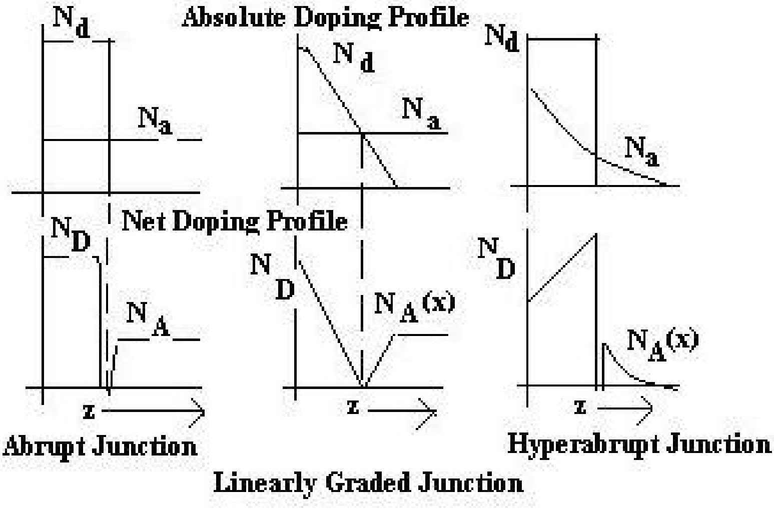


Figure 4. Doping profile of abrupt junction, linearly graded junction and hyper-abrupt junction.

$$N_A(z) = B\left(\frac{z}{z_0}\right)^m$$

z_0 = position of metallurgical junction along the longitudinal axis.

Depletion Width in a hyper-abrupt junction:

$d \sim d_p = K(V_R)^{1/(m+2)}$

$$C_{jD} = \frac{\varepsilon A}{d} = \frac{\varepsilon A}{K(V_R)^{\frac{1}{m+2}}}$$

In an abrupt junction or one sided step junction, m = 0 :

$$C_{jD} = \frac{\varepsilon A}{K(V_R)^{\frac{1}{2}}} \; : \; =$$

varies as 1/(Reverse Bias Voltage= V_R)1/2 ;

In linearly graded junction, m = 1:

$$C_D = \frac{\varepsilon A}{K(V_R)^{\frac{1}{1+2}}} = \frac{\varepsilon A}{K(V_R)^{\frac{1}{3}}} \; :$$

varies as 1/(Reverse Bias Voltage= V_{RR})1/3;

In hyper - abrupt junction, m = -

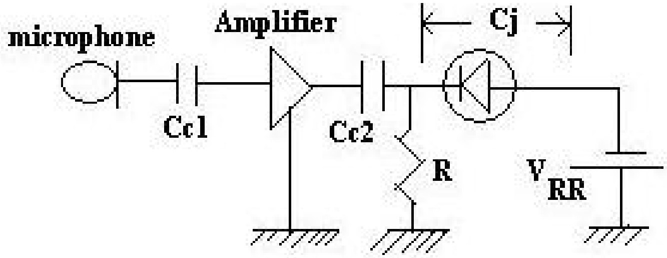
$$\frac{3}{2}$$

:

$$C_D = \frac{\varepsilon A}{K(V_R)^{\frac{1}{-(\frac{3}{2})+2}}} = \frac{\varepsilon A}{K(V_R)^2} =$$

varies as (Reverse Bias Voltage= V_{RR})2;

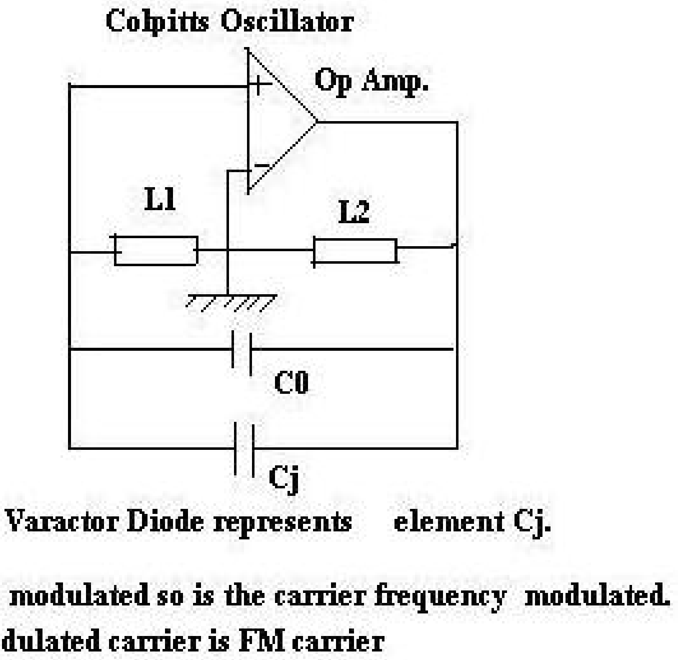
Hyper - abrupt Junction Diode is known as VARACTOR DIODE. It is used for Frequency Modulated Waves generation.



Arrangement for modulating a varactor diode capacitance Cj in accordance with a audio signal generated by a microphone.

This Cj can be used in RF oscillator to generate a FM carrier. The carrier frequency will be directly proportional to the reverse voltage Vj applied across Cj.

Figure 5. Arrangement for modulating Varactor diode junction capacitance by audio signal.



Varactor Diode represents element Cj.

As Cj is modulated so is the carrier frequency modulated. This modulated carrier is FM carrier

Figure 6. A Colpitts Oscillator with a varactor diode in its tank circuit.

The frequency of oscillation of Colpitts Oscillator is the Frequency of Resonance of the Tank Circuit.

Radial Frequency of Oscillation $= \omega_0 =$

$$\frac{1}{\sqrt{(L_1+L_2)(C_0+C_j)}};$$

Let $C_j/C_0 >> 1$ then $\omega_0 =$

$$\frac{1}{\sqrt{(L_1+L_2)(C_0+C_j)}}=\frac{1}{\sqrt{(L_1+L_2)\left(C_0)(1+\frac{C_j}{C_0}\right)}}=\frac{1}{\sqrt{(L_1+L_2)(C_0)}}\sqrt{\frac{C_0}{C_j}}\;;$$

Therefore ω_0 =

$$\frac{1}{\sqrt{(L_1+L_2)(C_0)}}\sqrt{\frac{C_0}{\frac{\varepsilon A}{K(V_R)^2}}}=\frac{1}{\sqrt{(L_1+L_2)(C_0)}}\sqrt{\frac{C_0}{\frac{\varepsilon A}{K}}}\times V_R$$

Thus we obtain frequency proportional to frequency. This generates FM carrier.

Physics of Diffusion Capacitance:

It can be shown that diffusion capacitance C_d can be obtained by the following relation:

For one sided step junction N+P,

$r_dC_d = \tau_n$ where τ_n life-time of minority carriers in lightly doped side.

In equally doped diode,

$1/ r_dC_d = 1/\tau_n + 1/\tau_p$ where τ_n life-time of minority carriers on P-type doped side.

τ_p life-time of minority carriers in N-type doped side.

In BJT in forward active mode:

$r_e C_d = \tau_t$

where r_e = incremental resistance of EB diode;

τ_t = transit time across the base.

SSPD_Chapter 3_Section3.3. Quantum Mechanical Diodes.
Section 3.3. Covers the Physics, Symbol and I-V characteristics of Zener Diode, Tunnel Diode and Backward Diode.

3.3.Quantum Mechanical Diodes.

Till now we have studied classical diodes. Classical diodes are fabricated using non-degenerate semiconductor. Doping of 10^{17}/cc or less is non-degenerate semi-conductor. Doping in the range of 10^{18}/cc or more gives rise to degenerate semi-conductor.

In non-degenerate semi-conductors we have drift and diffusion of carriers. In degenerate semi-conductor apart from drift and diffusion we have band-to-band tunneling which is a purely quantum-mechanical transport mechanism.

I have talked about Quantum Mechanical tunneling in Chapter 1, Part 8. We will briefly review it in this Section.

An electron in an infinite potential well can never come out of the potential well. Electron in finite potential well with thick potential wall can come out of the potential well if by photon excitation or by thermal excitation it gains sufficient energy to cross the surface potential barrier and escape into vacuum above 0 eV energy level. These processes are known as Photoionic-emission or Thermoionic-emission. But there is a third case where the electron is trapped in a potential well with thin potential walls. The potential walls are as thin as 10A° to 100A° as shown in the Figure ( ) in Chapter 1 Part8 .In this case there is a finite probability of finding the electron outside the potential well by the process of tunneling.

The Tunneling factor is calculated to be:

$$T = Exp\left[-\frac{2W\sqrt{2m_e(qV_1 - E)}}{\hbar}\right] \qquad 3.3.1$$

Substituting the values of the different parameters, the values of Tunneling Factor is tabulated in Table 3.3.1.

Table 3.3.1.Qunatum Mechanical Tunneling Ratio for different Wall thickness and different Barrier Potential.

W	$(qV_1\text{-}E)$	Exponent term	T	Comments

10A°	1eV		-10.2463		3.5×10^{-5}		With 10A° wall thickness and 1eV well depth there is a very low probability of finding 3.5 electron in 100,000 electrons.	
5A°	1eV		-5.12317		6×10^{-3}		With wall thickness reduced to 5A° but well depth kept constant tunneling increases to 6 out of 1000electrons.	
1A°	1eV		-1.02463		0.36		With wall thickness reduced to 1A° but	

				depth well kept constant tunneling increases to 36% .
1A°	10eV	-3.24018	0.04	With wall thickness kept at 1A ° but depth well increased to 10eV tunneling again decreases to 4% .
1A°	100eV	-10.2463	3.5×10^{-5}	With wall thickness kept at 1A ° but depth well increased to 100eV tunneling drastically decreases to . 3.5 electron

in
100,000
electrons.

As the doping density is increased we encounter 4 different classes of devices:

1. Classical Diode with an Avalanche Breakdown Voltage of BV_A from 30V to 100V.
2. Zener Diode with a zener breakdown of V_Z = 4V and below.
3. Backward Diode which conducts in forward as well as backwarddirections.
4. Tunnel Diode which has N-Type Negative Resistance Region(NIR).

The I-V characteristics of these four diodes are shown in Figure 3.19.

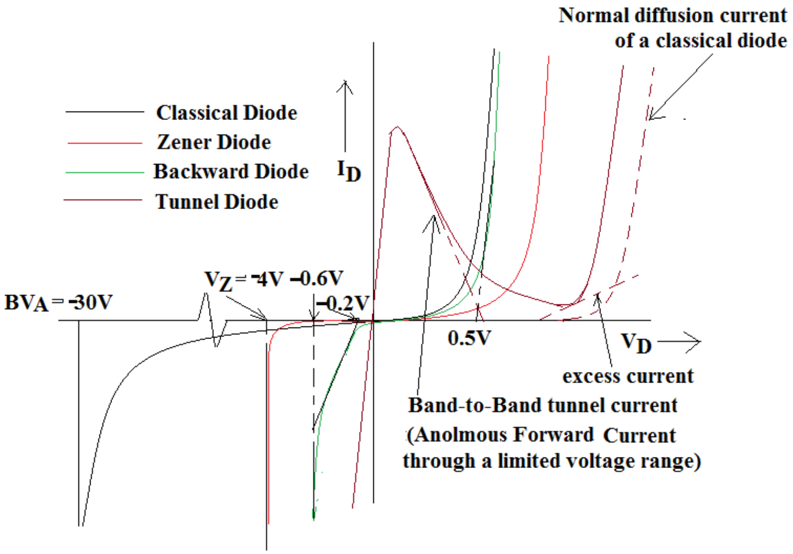


Figure 3.19. I_D-V_D characteristics of (a)Classical Diode(b)Zener Diode(c)Backward Diode and (d)Tunnel Diode.

We are considering an abrupt Si-diode with symmetrical N-type and P-type doping. Therefore the the built-in potential is given by the following formula:

$$\varphi_{BO} = V_{Th} \ln\left(\frac{N_A N_D}{n_i^2}\right) = 2V_{Th} \ln\left(\frac{N}{n_i}\right) \qquad 3.3.1.$$

Depletion Width (d) is given by the following fomula:

$$d_0 = \sqrt{\frac{2\epsilon}{q}\left(\frac{1}{N_A}+\frac{1}{N_D}\right)\varphi_{BO}} = \sqrt{\frac{4\epsilon}{q}\times\frac{\varphi_{BO}}{N}} \quad where\ \epsilon = \epsilon_0\epsilon_r$$

$$= \frac{10^{-12}F}{cm} \qquad 3.2.2$$

Using Equation (3.3.1) and (3.3.2), the built-in barrier potentials and the depletion widths are tabulated in Table 3.3.2.

Table 3.3.2. The built-in barier potential,depletion widths and Emax at zero bias for Classical Diodes(10ˆ16/cc), Zener Diodes(10ˆ20/cc) ,Back-ward Diodes(10ˆ21/cc) and Tunnel Diodes(10ˆ22/cc).

N	**10^{16}/cc**	**10^{19}/cc**	**10^{20}/cc**	**10^{21}/cc**	**10^{22}/cc**
Φ_{BO}	0.718V	1.077V	1.197V	1.317V	1.437V
d_0	4236.7A°	164A°	54.7A°	18.0A°	5.74A°
E_{max0}(V/cm)	33939.75	1.3×10^6	4.38×10^6	14.4×10^6	46×10^6

In Figure 3.20. I show that there is no possibility of Quantum Mechanical Tunneling.This is the usual classical case. In this under zero-bais, diffusion of majority carriers is exactly balanced by drift of minority carriers and detailed charge balance is maintained.Under forward bias, depletion width shrinks and potential barrier is reduced by forward bias hence the charge balance is heavily tipped in favour of majority carriers diffusing in the forward direction and it exponentially increases. Under reverse bias, diffusion of majority carriers is fully prevented because of higher potential barrier. Potential barrier has increased by reverse bias and only limited number of minority carriers are drifting down the potential gradient as reverse leakage current.Here the electrons from the top of the Valence Band cannot do Band-to-Band tunneling from Valence to Conduction Band. This is because the potential barriers thickness is thicker than 4236.7A° and Potential Barrier height is greater than 1V hence there is no possibility of tunneling asa is evident from Table 3.3.1.

In Forward Bias and Reverse Bias, thermal equilibrium is disturbed hence we donot have Fermi Level but we have IMREF levels and IMREF levels are not aligned.

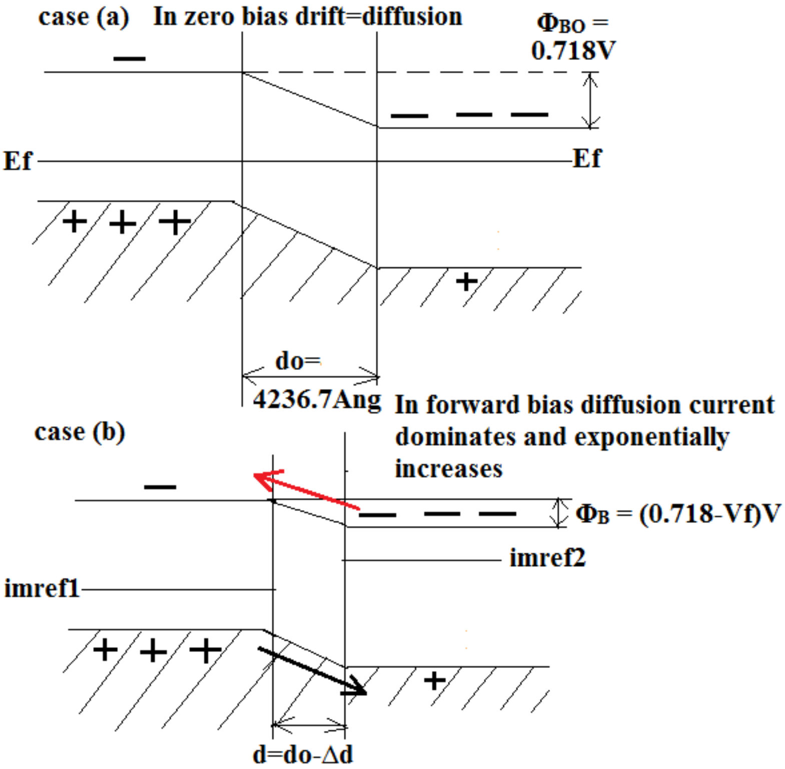

case (a)
In zero bias drift=diffusion
Φ_{BO} = 0.718V
Ef
Ef
do=
4236.7Ang
case (b)
In forward bias diffusion current dominates and exponentially increases
Φ_B = (0.718-Vf)V
imref2
imref1
d=do-Δd

Figure 20.A.(a)Zero-bias condition with do=4236 Angstrom and lower barrier potential = (0.718 - Vf) V;

(b) Forward-bias condition with narrower depletion width 'd' and lower barrier potential=(0.718-Vf)V;

case (c) Reverse-biased case. Only minority carriers are drifting down the potential hill as reverse leakage current.

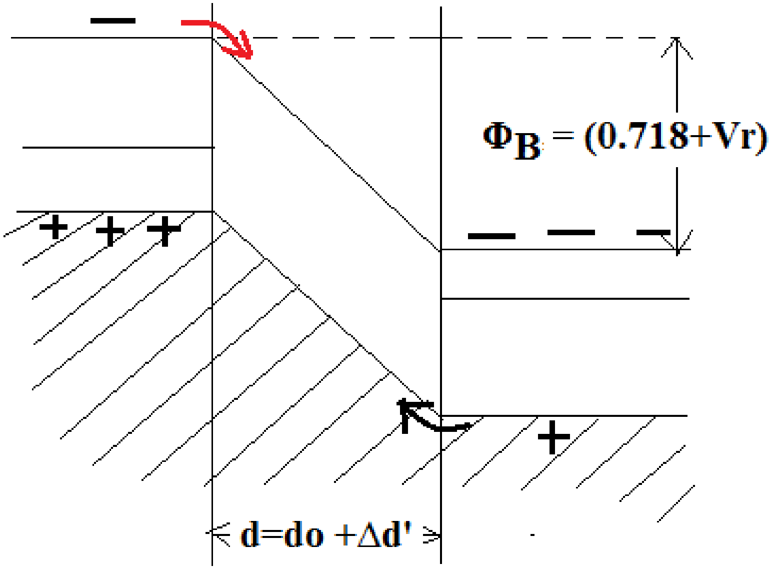


Figure 20.B.Reverse-biased diode with widened depletion width(do + Δd') and heightened barrier potential (0.718+Vr)V.

Now we examine the case of Zener diode with doping density of 10^{20}/cc. As seen from Table 3.3.2. the built –in barrier potential is 1.077V which means that Fermi level is very near the band edge in P-Type as well as in N-Type.The Potential Barrier Width is 164A°.

Figure 3.21 gives the Energy Band diagram of Zener Diode.Let us assume that its Zener Breakdown is 4V. As seen in the diagram, at reverse bias of 4V, the built-in Barrier Potential $= (1.077+4)\text{V} = 5.077\text{V}$ and the Barrier Width $=$ depletion Width $= 356\text{A}^\circ$. As seen from case (b) of Figure 3.21, Valence Band in P-Type faces empty permissible energy states across the barrier into the conduction band of N-Type.Since Barrier Width is thin and Barrier height is moderate, quantum mechanical tunneling occurs and Zener Breakdown occurs. In addition to the minority carriers drifting across the barrier the electrons from P-Side tunnel to the N-Side adding to the reverse current and leading to Zeber Breakdown.

When forward bias is done, forward current can flow only through diffusion across the lowered barrier potential.There is no possibility of tunneling.

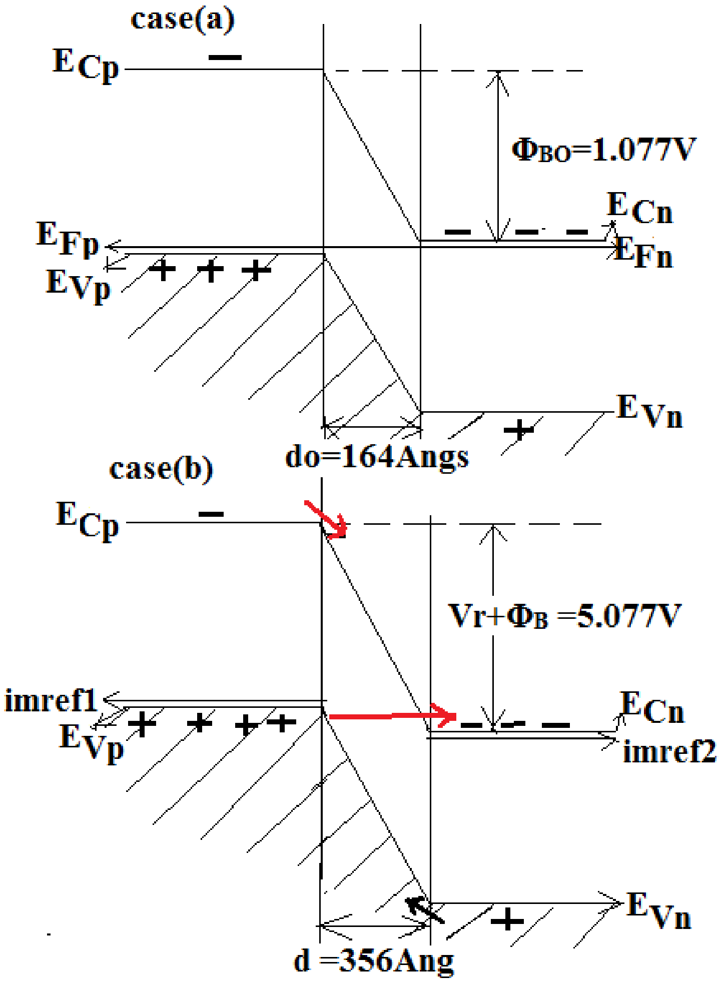


Figure 3.21.Energy Band Diagram of Zener Diode Case (a)Zero Bias, Case (b)Reverse Bias.

3.3.1. Design of Zener Diode for a given breakdown between 0V to 4V.

Experimentally it has been found that just as we have a critical field of 5×10^5V/cm for Avalanche Breakdown similarly we have a crtical field for

Zener Breakdown . This $E_{critical\_Zener} = 1.2 \times 10^6$ V/cm.

Hence if I am designing for a Breakdown Voltage of Vz Volts then the following Design Rule holds good:

$$\frac{E_{max} \times d}{2} = \frac{E_{crit} \times d}{2} = \frac{1.2 \times 10^6 \times d}{2} = total\ Potential\ Barrier\ Height = \varphi_{BO} + V_Z$$

Therefore:

$$d = \frac{2(\varphi_{BO} + V_Z)}{E_{crit}} \qquad 3.3.1.1.$$

From depletion width consideration:

$$d = \sqrt{\frac{4\epsilon}{q} \times \frac{(\varphi_{BO} + V_Z)}{N}} \qquad for\ symmetrically\ doped\ diode \qquad 3.3.1.2.$$

Built-in Barrier Potential:

$$\varphi_{BO} = 2V_{Th} \ln\left(\frac{N}{n_i}\right) \qquad 3.3.1.3$$

Combining the Equations (3.3.1.1),(3.3.1.2) and (3.3.1.3) we get;

$$N\left(V_Z + 2V_{Th} \ln\left(\frac{N}{n_i}\right)\right) = \frac{\epsilon_0 \epsilon_r E_{crit}^2}{q} = 9.47 \times \frac{10^{20} V}{cm^3} \qquad 3.3.1.4$$

Using the Design Formula (3.3.1.4) we tabulate the Fabrication Parameters of Zener Diode of Vz = 4V,3V,2V and 1V in Table 3.3.1.1.

Table 3.3.1.1.Doping Density for Zener Diode of Zener Breakdown Voltage = 4V,3V.2V and 1V.

V_Z	N_A(No.per cc)	N_D(No. per cc)
4V	1.81×10^{20}	1.81×10^{20}
3V	2.235×10^{20}	2.235×10^{20}
2V	2.91×10^{20}	2.91×10^{20}
-1V	4.18×10^{20}	4.18×10^{20}

3.3.2. Comparison between Zener Break-down Voltage and Avalanche Break-down Voltage.

The following Table brings out the major differences between two kinds of Breadowns.

Table 3.3.2.1. Salient features of Avalanche Breakdown and Zener Breakdown.

	Avalanche Breakdown	**Zener Breakdown**
$E_{critical}$	5×10^5V/cm	1.2×10^6V/cm
Mechanism	Impact collision and ionization	Covalents are strained and broken
Voltage Range	Above 6V	Below 4V
Between 4V and 6V the mechanism is partially		

Avalanche and partially Zener.		
Sharpness of Breakdown	Soft BreakdownSlope resistance 40Ω	Sharp BreakdownSlope Resistance lower than 10Ω
Temperature Coefficient	Positive Temperature Coefficient	Negative Temperature Coefficient

It is possible to design a Zener Diode between 4V and 6V which has Zero Temperature Coefficient. This will act as an ideal Voltasge Source as well as an ideal Reference Voltage.

Chapter 3_Section3.3.3Tunnel Diodes,3.3.4.Backward Diodes

Chapter 3_Section 3.3.3 and Section 3.3.4. conclude the discussion on Quantum Mechanical Diodes by giving the Physics of Tunnel Diodes and Backward Diodes.

3.3.3. Tunnel Diode.

Tunnel Diode is heavily doped and experiences tunneling in both forward and backward directions as shown in Figure 3.22. Degenerate doping causes high densities of crystalline imperfections through out the bulk. These introduce intergap states. These states introduce new conduction mechanism in forward bias for example carrier tunneling to and from such states within the transition region. Hence there is an excess current component. The normal diode characteristics in injection region donot correspond to that of the tunnel diode.Because of heavy doping the Fermi-level penetrates both Conduction Band on N-Side and Valence Band on P-Side.

The built-in barrier Potential is larger than Band-Gap Voltage of 1.12V. Built in Poential is of the order of 1.3V hence at 0.1V forward bias there is no notiveable forward current by diffusion but there is forward current due to tunneling of electron from N-Side to the empty energy states on P-Side as seen in Figure 3.22.B shown by the blue arrow. Hence there is a distinct band-to band tunneling current in forward direction.

As seen in Figure 3.22.C. there is maximum alignment for tunneling from N-Side to P-Side.This results in maximum forward tunnel current as shown in I-V graph of Figure 3.22.C.

With forward bias around 0.3V, the tunneling from N-Side to P-Side starts getting misaligned and tunneling current starts decreasing but diffusion across the barrier is still not starting as shown in Figure 3.22.D.

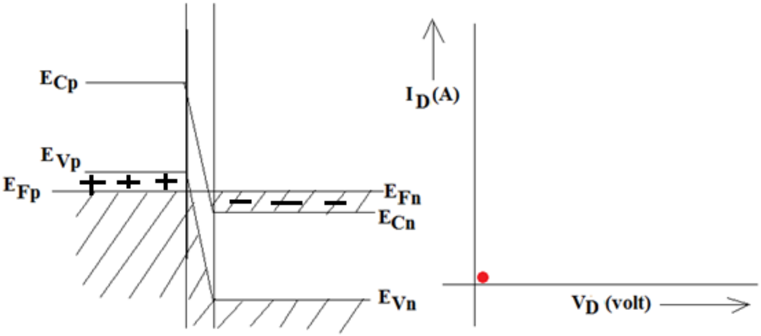


Figure 3.22.A. Tunnel Diode at Zero Bias. There is no current flow.

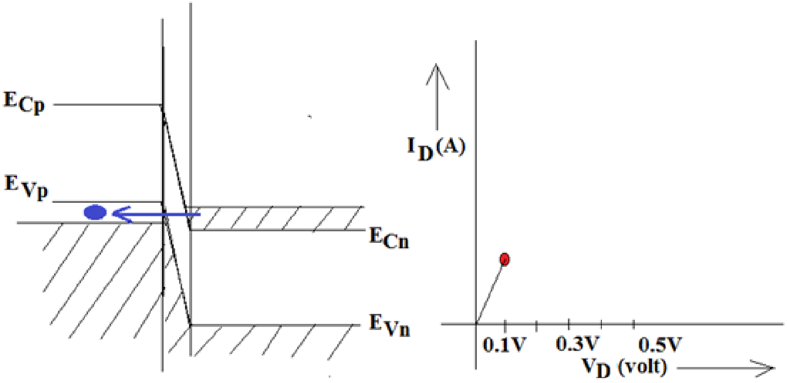


Figure 3.22.B. Under Forward Bias of 0.1V no noticeable diffusion current but distinct babd-to-band tunneling current as shown by the blue arrow.

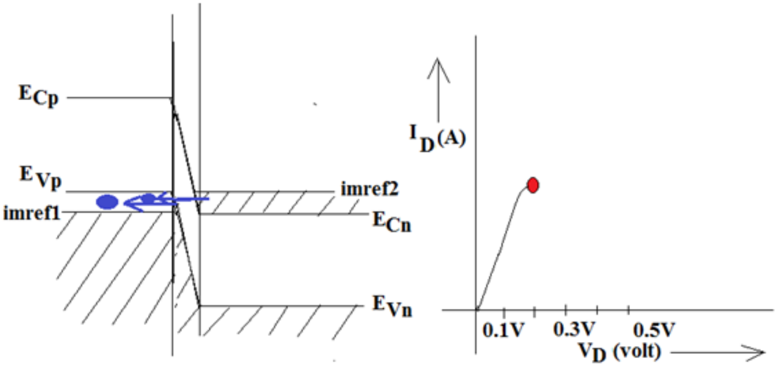


Figure 3.22.C.Maximum alignment between the occupied band above Ec (in the N-Side) and vacant band under Ev(in the P-Side).This results in maximum tunnel current.

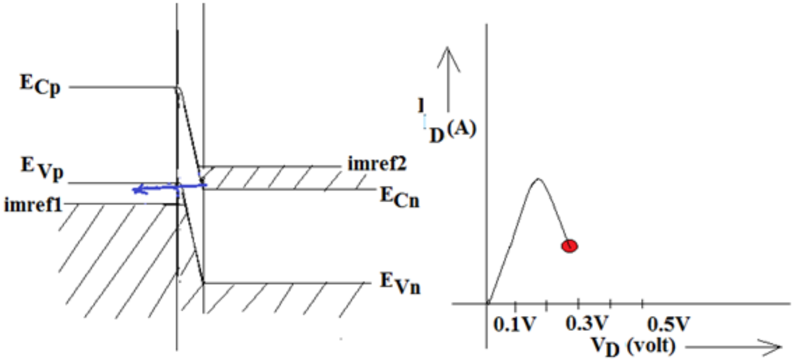


Figure 3.22.D. With greater forward bias around 0.3V the band-to-band tunneling starts getting mis-aligned and tunneling current decreases as shown in the graph but the potential barier is not low enough to cause diffusion across the barrier.

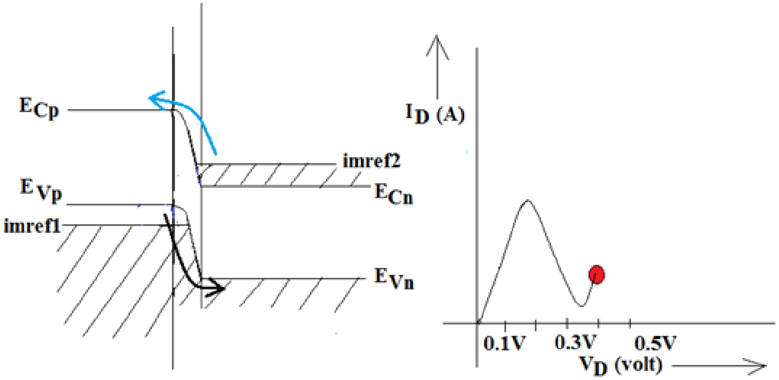


Figure 3.22.E. With further increase in forward bias, tunneling stops and diffusion of majority carriers across the much reduced potential barrier starts.

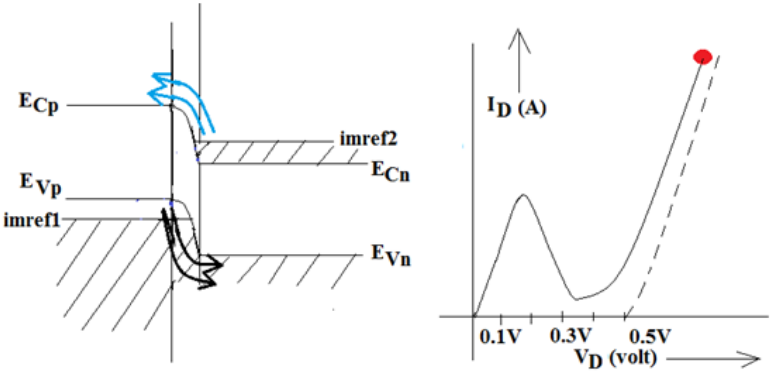


Figure 3.22.F.With forward bias greater than 0.5V(the knee voltage of Si-Diode)the classical diffusion of majority carriers is restored aross the barrier.

Under reverse bias, Zener Breakdown immediately occurs as the Emax at the mettalurgical junction exceeds the critical field as shown in Table 3.3.2. Hence total I-V characteristic is as shown in Figure 3.22.G.

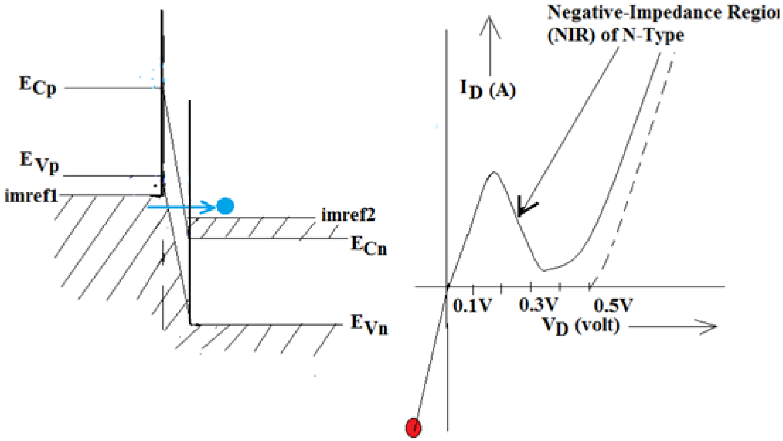


Figure 3.22.G.Under Reverse Bias at small reverse voltage Zener Breakdown occurs and large current in reverse direction starts flowing giving it a backward diode characteristics in backward direction and tunnel diode characteristics in forward direction.

As seen in Figure 3.22.G. there are two features of

Tunnel Diode:

1. It is highly non-linear ;
2. And it has a strong N-Type Negative-Impedance Region(NIR) characteristics.

The I-V characteristics of Tunnel Diode is parameterized as shown in Figure 3.23.

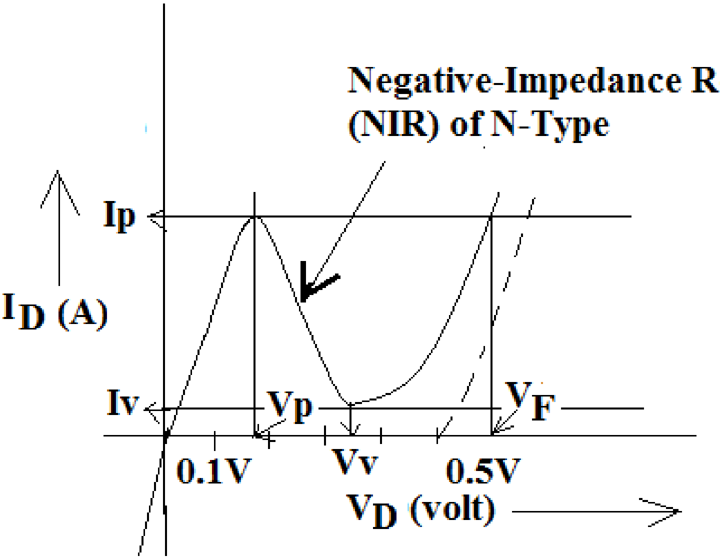


Figure 3.23.I-V characteristics of a Tunnel Diode and its D.C.parameters.

A Tunnel Diode exhibits Negative Inmpedance Region for GaAs as well as for Si and Ge material.The typical DC parameters of Si,Ge and GaAS are tabulated in Table 3.3.3.1.

Table 3.3.3.1. Typical DC parameters for Si,Ge and GaAs Tunnel Diodes.

	Si	**Ge**	**GaAs**
I_p/I_v	3.5	8	15
V_p(V)	0.065	0.055	0.15
V_v(V)	0.42	0.35	0.5
V_F(V)	0.7	0.5	1.1

It is evident from the Table that GaAs Tunnel Diode has best D.C. parameters leading to strongest NIR characteristics and hence best suited for Microwave Oscillators.

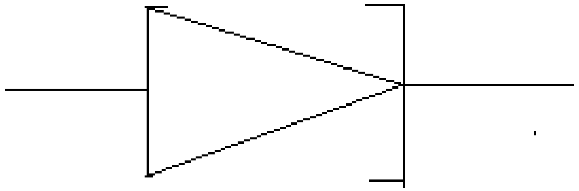
Table 3.3.3.2 gives the relative merits and demerits of the Tunnel Diode.

Table 3.3.3.2. Merits and demerits of Tunnel Diode.

Merits	**Demerits**
Low cost	Two terminal decvice hence no isolation between I/P and O/P.
Low noise	
Simple	
High Speed	
Environmental immunity	
Low power consumption	

Device Symbol:

Tunnel Diode Symbol



Its incremental equivalent circuit is as follows:

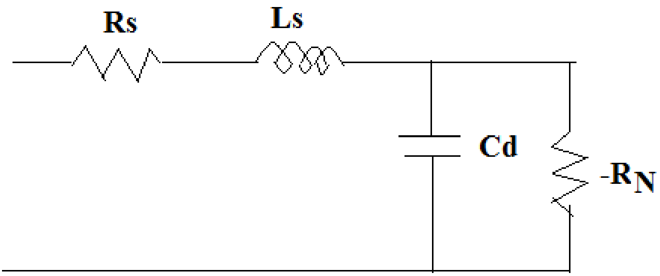


Figure 3.24.Small Signal Model of Tunnel Diode 1N2934

In the above incremental circuit following are the incremental parameters:

R_S = series resistance = 1 Ω = ohmic resistance due to leads, diode bulk and diode contact.

L_S = series inductance due to terminal leads at very high frequency = 5nH.

C_d = diffusion Capacitance under forward bias

= 20pF.

$-R_N = -30\ \Omega$.

Because of Quantum Mechanical Tunneling, the switching speed is at the speed of light. Only C_j[Junction Capacitance = (ϵA/d)] and C_d[Diffusion Capacitance where $C_d r_d = \tau_{eff}$(effective life-time) which is very short due to degenerate doping] limit the switching speed.Typical Switching Delay is 50ps therefore switching speed of 20GHz has been achieved. It is being used as microwave oscillator in hand-sets of mobile telephony.

3.3.4.Backward Diode.

Backward diode is like a tunnel diode with asymmetrical doping so that in Forward Bias, tunneling is suppressed and it behaves like a classical diode and in reverse bias it behaves like a Zener diode with Zener Breakdown at 0Volt.

The I-V characteristics is shown in Figure 3.25:

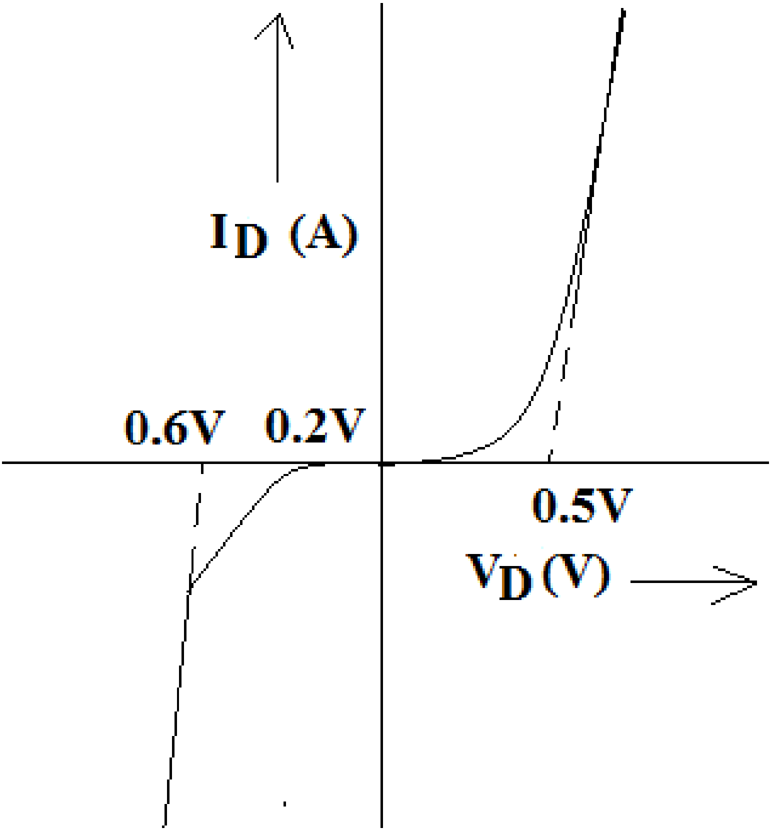
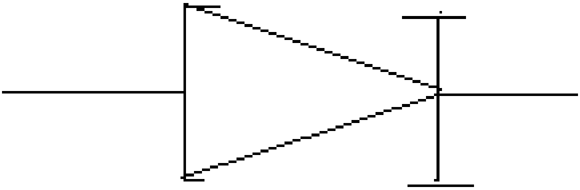


Figure 3.25. I-V charateristics of a Backward Diode.

Symbol of Backward Diode



SSPD_Chapter 3_Section 3.4.Schottky Diode and Section 3.5. Ohmic Contact.
Section 3.4 describes the physics and application of Schottky Diode and Section 3.5. describes the procedure for realizing ohmic contact.

3.4.Schottky Diode.

Metal Aluminum is the commonly used interconnection path in Integrated Circuit Technology because it provides a good conducting path and it makes ohmic contacts on the contact pads of the solid state devices being integrated. Bulk 'Al' has a resistivity of $2.7\times10^{-6}\Omega$-cm but in thin film form its resistivity is higher by a factor of 20 hence it high current densities as high as ~ 10^5A/cm^2 which is required for high speed switching networks it leads to serious reliability problems of a phenomena known as electromigration which is a major cause of the breakdown of VLSI and ULSI integrated circuits.

At nodes (length of the channel) below 90nm, the pitch (the lateral width of the interconnect) of interconnection is 45nm or less. At this level of integration, to build complex and dense circuits the multilevel metallization is routinely employed. As much as 10 multi layers are used. The metal thickness ranges from a fraction of µm to several μm. The thinner interconnects route signals while thicker layers serve as power bus. Electrical connection between two adjacent levels of interconnects are made through a VIA. To reduce the contact resistance at a contact pad a Silicide such as $NiSi_2$ is added. In electromigration (also known as the electron wind effect) metal ions are forced to move downstream due to the high electron current densities, along the the grain structure. This results into voids and hillocks in the film at curves and forks.The thinning due to voids leads to excessive heating and break in the conducting path.

Electromigration is prevented by controlling the grain structure along the microcrystallites that form the metal lines. Larger grains have less surface area and therefore resist electromigration. Some alloying metals, such as copper, accumulate in the grain boundaries, locking the atoms thetre into place and thus preventing electromigration. Copper also prevents hillocking of Al films and thus prevents non-unifporm thermal effects. It is thus increasingly common to add Cu to Al in small quantities to form interconnects. Since 1998 IBM has introduced Cu interconnects leading to a much higher performing ICs with no problem of electromigration even at interconnect pitch lower than 90nm. Copper has replaced Aluminum as the interconnect material in nodes below 90nm. Cu has excellent electromigration reliability and 40% lower resistance than Al. It may be deposited by electroplating or by CVD. Copper patterns are etched by damascene process. This wil be given in supplementary sheets.

But ‘Al’ is itself P-Type material as is evident from Figure 3.26 and Table 3.4.1. Hence its deposition on P-Type Semi-conductor will give an ohmic contact but its deposition on N-Type Semiconductor gives rise to rectifying contact and such an interface is a PN Junction diode which in reality is Metal-Semiconductor Diode with very fast switching properties. This Metal Semiconductor was studied by Schottky in great details and hence it is called Schottky Diode.

In 1874, Braun noted the dependency of total resistance on the polarity of the applied voltage and on the surface conditions. By 1904, point contact diodes started coming into practical use. These were known as ‘Cat Whiskers’ used in crystal diode in demodulating the AM Radio Broadcast Signals. In 1931, Wilson gave the transport theory based on the Band Theory of Solids. In 1938, Schottky suggested that contact potential could develop at metal-semiconductor interface. The model based on this is known as Schottky Model. In 1938, Mott gave the model for swept-out Metal-Semiconductor contact. This is known as Mott barrier. In 1957, Henish in his book ‘Rectifying Semiconductor Contacts’ summarized the basic theory and the historical development of rectifying Metal-Semiconductor

Contacts.

Table.3.4.1.Position of III and V elements energy states within the band-gap of Silicon.

The energy values refer to the energy below conduction bandedge in case of donor and refer to the energy above valence bandedge in case of acceptor.

[Reference: ‘Semiconductor Devices-basic principles’, by Jasprit Singh, Chapter 2,pp 74, 2007 reprint, Wiley –India Publication]

Semiconductor	**Impurity(Donor)**	**Ed(meV)**	**Impurity(Acceptor)**	**Ea(meV)**
Si	Li	33	B	45
	Sb	39	Al	67
	P	45	Ga	72
	As	54	In	100
	Mn(deep EnergyStates)	530	Au(deep EnergyStates)	540
			Zn(deep EnergyStates)	550

Deep energy states in the band-gap act as the effective stepping stone for recombination hence they act as the Life-Time killer. Their introduction will lead to reduction in minority carriers storage under forward bias condition.Hence in 70s to reduce the storage delay time, Gold was introduced as the life-time killer. But subsequently this was discarded and Schottky Clamped transistors were used to reduce the minority carrier storage delay. Nowadays Schottky-clamped TTL gates have become the norm of the day. These are available as 74AS series.

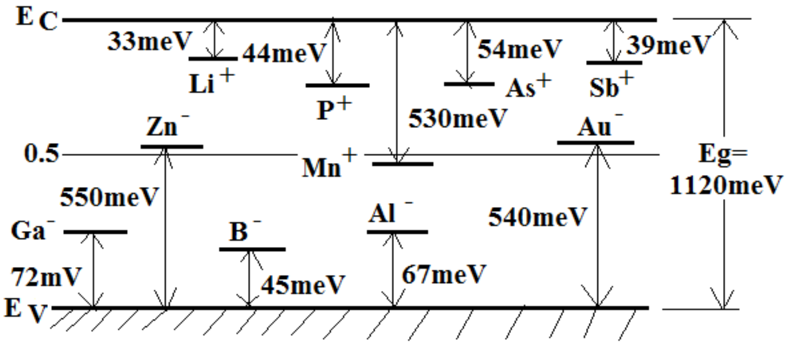


Figure 3.26. The introduction of shallow energy states by most of the Donor(Gr.V) and Acceptor(Gr.III) Dopents except Zinc, Manganese and Gold. Zinc, Manganese and Gold introduce deep energy states hence they greatly enhance recombination therefore these impurities are known as Life-Time Killer.

3.4.1. Physics of Schottky Diode.

Aluminum layer on N-Type Semiconductor forms a rectifying junction with a built-in barrier potential as shown in Figure 3.27. and Figure 3.28.

In Figure 3.27. Case (a):

Work-function of Metal = $q\Phi_m$ (eV).

Work-function of Siliconl = $q\Phi_{Si}$ (eV).

Electron affinity in Silicon = χ_{Si} (eV).

Barrier Potential Energy seen from metal side = $q\Phi_B = q(\Phi_m - \chi_{Si})$.

Barrier Potential seen from semi-conductor side = $V_{bi} = (\Phi_m - \Phi_{Si})$.

Built-in Barrier Potential on the semi-conductor side can be expressed as = $V_{bi} = (\Phi_B - \Phi_D)$.

These are in ideal condition. In real – life, there are surface states at the interface which complicate the situation and measured values of Φ_B are much greater than the theoretical values. Table 3.4.1. gives the measured Schottky Barrier heights for electrons on N-type Silicon $\Phi_{B\square N}$ and the measured Schottky Barrier heights for holes on P-type Silicon $\Phi_{B\square P}$

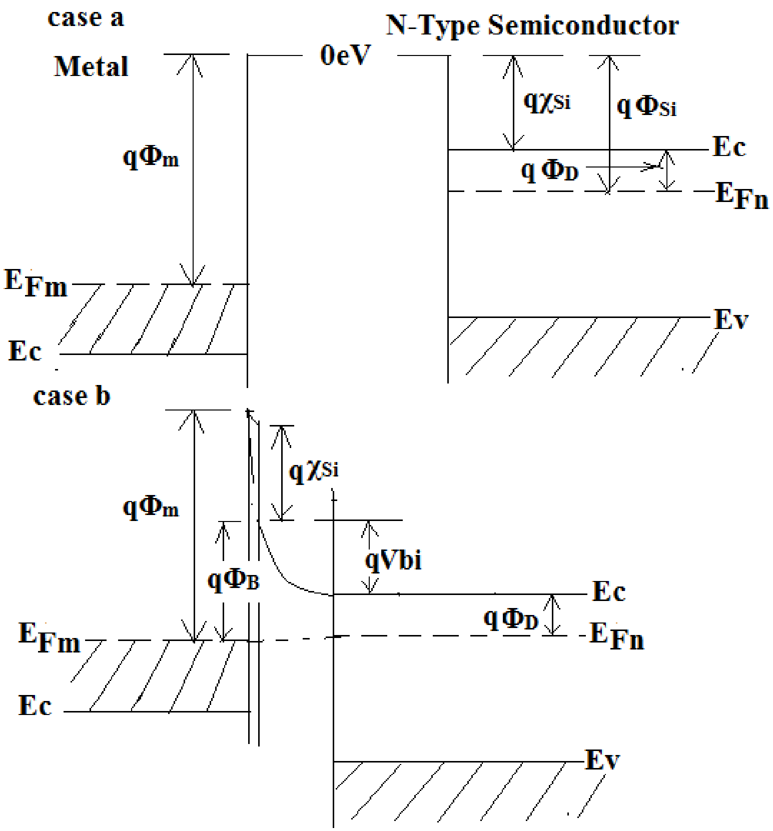


Figure 3.27. Energy-Band Diagram of Metal and Semiconductor (case a)when well apart and (case b) when in intimate contact.

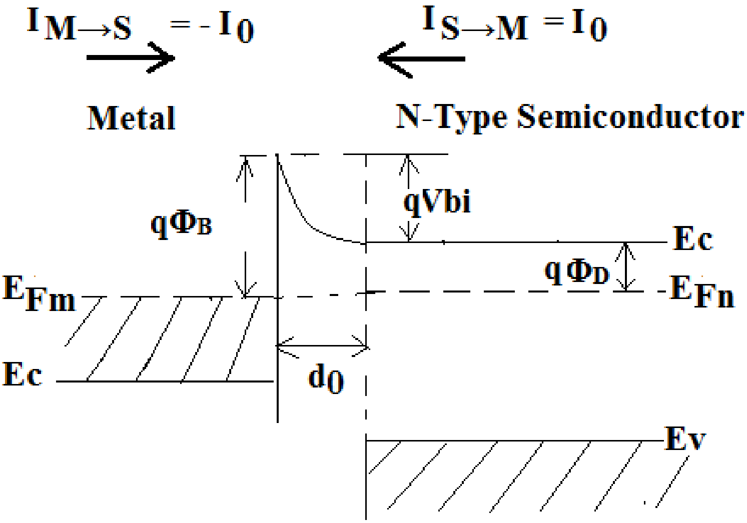


Figure 3.28. At Metal-Semiconductor interface there is potential barrier. Electrons on Metal side see a potential-barrier of Φ_B Volt and electrons on N-Type semi-conductor sees a potential barrier of Vbi.

Table 3.4.1. Measured Schottky Barrier heights for electrons on N-type Silicon Φ B □ N and the measured Schottky Barrier heights for holes on P-type Silicon Φ B □ P .

Metal	Mg	Ti	Cr	W	Mo	Pd	Au	Pt
Φ_M(eV)	3.7	4.3	4.5	4.6	4.6	5.1	5.1	5.7
Φ_{B}□N(eV)	0.4	0.5	0.61	0.67	0.68	0.77	0.8	0.9

Φ_{B}N(eV)	0.61	0.50	-	0.42	-	0.3	-

Under no bias, there is equal current flowing across the potential barrier from two sides. Hence there is detailed charge balance is maintained and net current across the Schottky Barrier is zero.

The Barrier Potential seen from the metal side is Φ_B and remains unchanged even under bias condition. Hence current from Metal to Semiconductor is kept at $-I_0$ under forward as well as reverse biased condition..

But the Barrier Potential from the semiconductor side is (V_{bi} -V_F) under forward bias condition just as in Classical Diode and as shown in Figure 3.29. Hence carrier diffuses across the forward biased junction with a greater ease and the current from Semiconductor to Metal is I_0Exp[V_F/V_{Th}]. Therefore net current across the junction under forward bias condition is:

$$I_D = I_{D0}\left(Exp\left[\frac{V_F}{V_{Th}}\right] - 1\right) \qquad 3.4.1.$$

Under reverse bias condition, the barrier potential from Metal-side is unchanged but from the Semiconductor-side it is (V_{bi} + V_F) as shown in Figure 3.30. Hence electron diffusion across the enhanced potential barrier is completely stopped and net current across the junction is:

$$I_D = -I_{D0} \qquad 3.4.2.$$

Equation 3.4.1 and Equation 3.4.2 show that Schotky diode behaves exactly like a classical Diode except for a lower knee Voltage at 0.1V because of a reduced barrier potential.

Detailed analysis of Schittky Diode will be done in Supplementary Sheets.

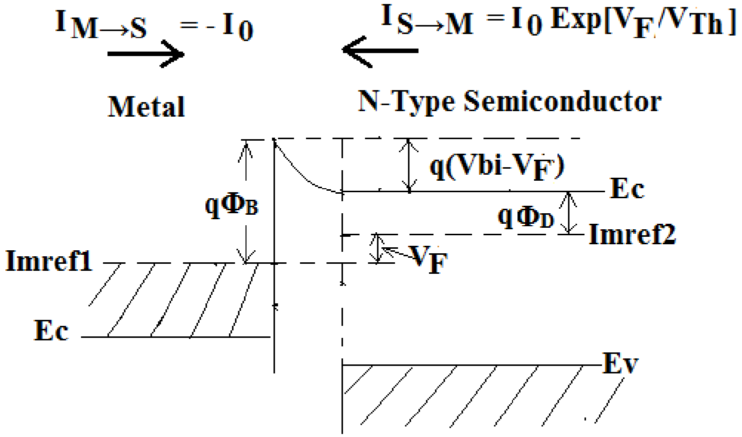


Figure 3.29. Under Forward Bias, barrier potential from semi-conductor side is reduced hence current flow from semi-conductor to metal side exponentially increases just as in Classical Diode.

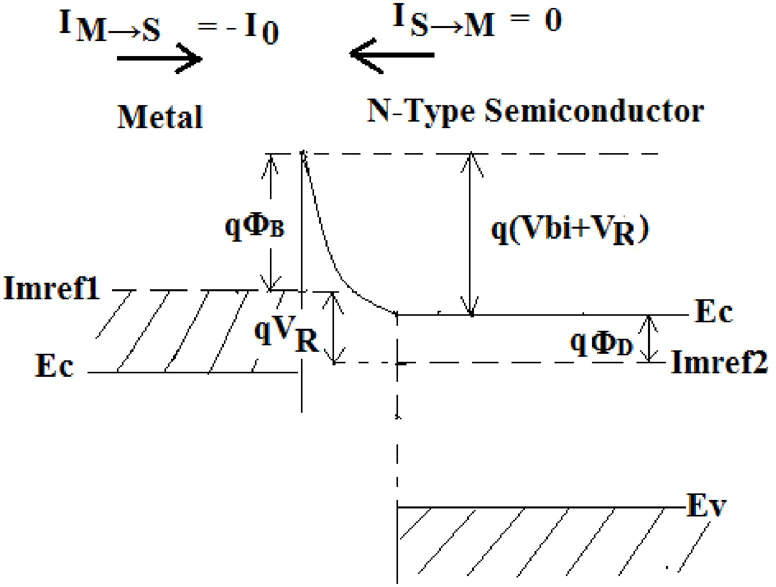


Figure 3.30. Under reverse bias, barrier height from Metal-side is unchanged but from Semiconductor side it is enhanced hence current flow from semi-conductor side is zero.

3.4.2. Data-Sheet of Schottky Diode.

As shown in Section 3.4.1, there is no minority carrier involved in the diode action. It is electron bothways. Hence there is no minority carrier storage and hence as we will see in the switching transients we are never confronted with mionority carrier storage delay in switching transients. Therefore Schottky Diode is known as ‘Hot-Electron Diode’.

It has a lower cut-in voltage of 0.1V because of lower barrier potential. And it has much lower junction plus diffusion capacitance. In Table 3.4.2.1. comparative data sheets of a Schottky Diode 1N5711 and Classical Diode 1N4001 are given.

Low fotward voltage drop and fast switching makes it ideal for protection of MOS devices, steering, biasing and coupling diodes for fast switching and low logic level application. It has higher cut-off frequency, it is reproducible and it has ease of fabrication.

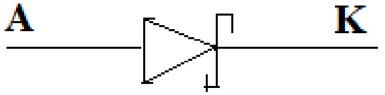
In Figure 3.31, its symbol and I-V characteristics in comparison to I-V characteristics of a classical diode is shown.

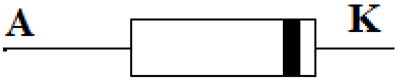
Table 3.4.2.1. Comparative Data Sheets of Schottky Diode 1N5711 and Classical Diode 1N4001.

1N5711(Schottky Diode)			**1N4001(Classical Diode)**		
PIV		70V	PIV		50V
P_D		400mW	P_D		500mW
I_{D0}		200nA At $V_R = 50V$	I_{D0}		5µA At $V_R = 50V$
V_γ	0.1V	At 10µA	V_γ	0.5V	At 10µA

V_F	0.4V	At 1mA	V_F	0.65V	At 1mA
C_{Tot}		2pF			20pF
t_{rr}(ReverseRecovery time)		1ns			3μs

Symbol of Schottky Diode





Black Band is on Cathode-side.

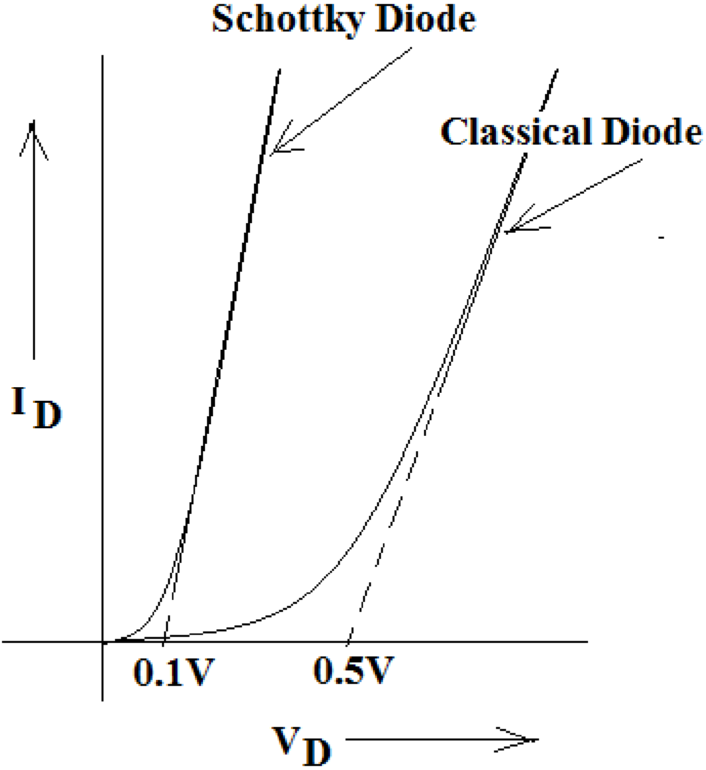


Figure 3.31.I-V characteristics of Schottky diode as compared to that of a classical diode.

3.4.3. Application of Schottky Diode in TTL Gates.

TTL Gate (Transistor-Transistor Logic)74Series is the most commonly used Saturating Logic Family. It suffered from a serious drawback and that was of SPEED. TTL under HIGH conditiont is in cut-off region and in LOW condition it is in deep-saturation. While switching from LOW to HIGH, it suffered from considerable storage delay because of deep-saturation and because of excessive minority carrier storage in the BASE in saturation mode. But if we put a Schottky Diode in parallel to BC Juncttion then even before BC junction is fully forward biased Schottky Diode is turned on.This prevents the BJT from going into deep saturation. A Schottky Clamped Transistor as shown in Fig 3.32. is barely saturated when it goes to LOW condition. Therefore it quickly comes out of Saturation Region and moves to Cut-Off region.This considerably improves the switching speed and toggle rate. The cross-section and the symbol of a Schottky Clamped Transistor used in TTL gate is shown in Figure 3.32. Such a TTL gate is known as Schottky Clamped TTL in 74AS series.

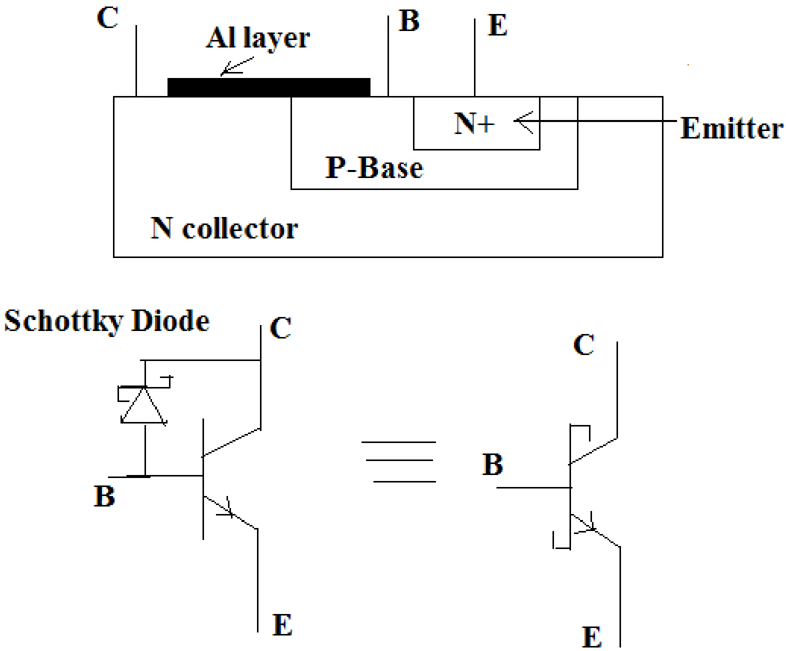


Figure 3.32. Fabrication of Schottky-Clamped Transistor and symbol of a Schottky Clamped Transistor.

3.5. Ohmic Contact.

Aluminum contact pad to P-Type semiconductor is an ohmic contact but Aluminum contact pad to N-Type semiconductor is a rectifying contact which we have seen to be Schottky Diode as seen in Figure 3.32. In Si-ICs, ohmic contact to N-Type Collector is never made directly. First a N+ contact region is diffused in N Collector and then Al contact pad is deposited on N+ contact region as shown in Figure 3.33.

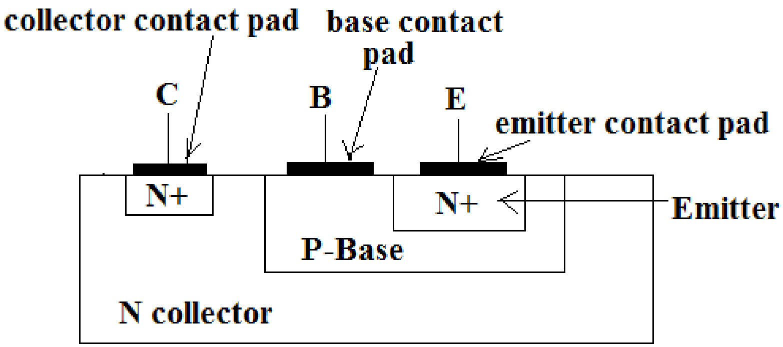


Figure 3.33.Cross-sectional area of a Vertical NPN Transistor in IC Technology.

In nanotechnology era, greater care is taken in making contact with N-semiconductor as show in Figure 3.34.

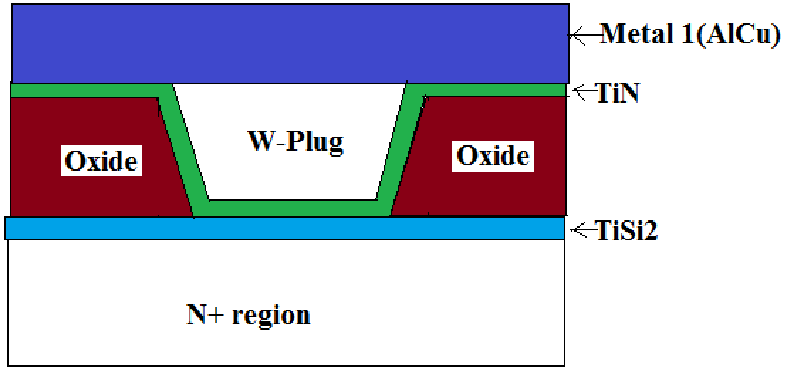


Figure 3.34. Contact Pad to N-Region for realizing a Ohmic-Contact.

As seen in Figure 3.34. upper layer of heavily doped N+ region is converted into Titanium Silicide. Over this Silicide layer, Oxide layer is deposited. A contact hole is made in Oxide Layer to reach

Titanium Silicide. A thin layer of Titamium Nitride is deposited lining the contact hole. This prevents the reaction and inter-diffusion between Silicide and Tungsten Plug.Tungsten Plug is deposited by CVD to fill the contact hole. Over the Tungsten Plug a layer of Metal(Al with some minute quantity of Cu) is deposited to make contact with the underlying N+ region via the tungsten Plug. This gives an ideal ohmic contact.

Tutorial Sheet of Chapter 3.Diode Physics.
This is a tutorial sheet on Chapter 3 on Diode Physics.

Tutorial Sheet of Chapter 3.Diode Physics.

Question 1. Pspice is a software which is used for Circuit and System Simulation. The Diode adopted in it, if not specified, has the following fabrication parameters and geometrical dimensions:

P-doping	**10^{16}/cc**	**N-doping**	**5×10^{15}/cc**	**comments**
Abrupt Junction	m = 1/2	NA	NA	C_j varies as $1/\sqrt{(\Phi_{BO} + V_R)}$
W_P	5μm	W_N	5μm	Bulk Width on two sides
Cross-SectionalArea	NA	NA	NA	100μm × 100μm
Diffusion Coeff.	$D_P = 10cm^2/s$	NA	$D_N = 30cm^2/s$	NA
Life Time	$\tau_P = 0.1\mu s$	NA	$\tau_N = 0.1\mu s$	τ(depletion) = 0.
Mobility	$384cm^2/$	NA	$1153cm^2/$	NA

	(V s)	(V s)
Diffusion Length	0.001cm = 10μm	0.001732cm = 17.32μm

NA-not applicable.

i.The current density should not exceed 10^3A/cm^2.If the current density exceeds this value it is going to disrupt Aluminum interconnections. Nowadays with the scaling reaching its limit, the above current density is being exceeded leading to the failure of the devices. Hence Al interconnection is being superseded by Cu interconnection.

Hence if the maximum rated current is 100mA then the cross-sectional area should be 100μm × 100μm.

ii.Since Diffusion Length is longer than Bulk width hence it is Narrow Bulk Diode.

iii.The total depletion width =

$$d_0 = \sqrt{\frac{2\varepsilon}{q}\left(\frac{1}{N_A}+\frac{1}{N_D}\right)\emptyset_{B0}} = 0.5\mu m \text{ where } \emptyset_{B0} = V_{Th}\ln\left(\frac{N_A N_D}{n_i^2}\right)$$

$= 0.68$V

$$d_p = 0.167\mu m \text{ and } d_n = 0.333\mu m$$

iv.Junction Capacitance =

$$C_{jo} = \frac{\epsilon A}{d_0} = 2pF$$

;

v. Reverse Saturation Current = Reverse Leakage Current for Narrow Bulk Diode =

$$I_{D0} = qA\left[\frac{D_p}{W_N - d_n} p_n^{-} + \frac{D_n}{W_p - d_p} n_p^{-}\right] = 10fA$$

vi. Reverse Leakage Current for a Wide Bulk Diode (when Bulk Width is 10 times the diffusion length then we fulfill the criteria for Wide Bulk Diode) =

$$I_{D0} = qA\left[\frac{D_p}{L_p} p_n^{-} + \frac{D_n}{L_n} n_p^{-}\right] = 6fA$$

As we see the leakage current in Narrow Bulk Diode is larger than that in Wide Bulk Diode which is not desirable but its switching transients are much better as the minority carrier stored can be reduced considerably by reducing the bulk width.

vii. The Calculation of Avalanche Break-Down Voltage = BV_A :

At the Avalanche breakdown point, E_{max} at the metallurgical junction is equal to $E_{critical}$ = 5×10^5 V/cm and let the depletion width be d* then from the first principles:

$$\frac{d^* \times E_{critical}}{2} = \emptyset_{BO} + BV_A$$

Solving this equation we get: $\sqrt{}(\Phi_{BO}+BV_A) = 15.2989$

Therefore: $(\Phi_{BO}+BV_A) = 234.05$

Therefore $BV_A = 233.37$V.

Calculate : i. Reverse Saturation Current [Answer 10^{-14}A = 10fA]

ii.Junction Capacitance at zero bias [Answer 2pF]

iii.Built-in barrier Potential [Answer 0.68V]

iv.BV_A (Avalanche Breakdown Voltage) [Answer 233V]

Once the design is complete the fabrication parameters, the method of diffusion and geometrical dimensions are sent to the Solid State foundry where the given Diode is fabricated and sent back to the customer for testing the design parameters.

Question 2. For an abrupt junction diode, $N_A = 10^{17}$/cc and $N_D = 10^{16}$/cc.

1. Determine built-in barrier potential?[0.778V]
2. Determine total depletion width under no bias condition?[0.33μm, here permittivity

$\varepsilon = 10^{-10}$F/m]

Question 3. Fill up the Table Below:

N_A(per cc	N_D(per cc	Φ_{BO}(V)	d_0(μm)
10^{17}/cc	10^{16}/cc	?	?
10^{19}/cc	10^{16}/cc	?	?
2.5×10^{16}/cc	10^{16}/cc	?	?

Answer:

N_A(per cc	N_D(per cc	Φ_{BO}(V)	d_0(μm)
10^{17}/cc	10^{16}/cc	0.778	0.327
10^{19}/cc	10^{16}/cc	0.898	0.335
2.5×10^{16}/cc	10^{16}/cc	0.682	0.345

Question 3. Determine E_{max} and d_{no} and d_{po} in Problem 2?

[E_{max} = 47.584KV/cm, d_{po} = 0.0297μm, d_{no} = 0.297μm]

Question 4. If reverse saturation current = 40fA then determine I_D for V_D = 0.7V and V_D = 0.55V[Answer 19.7mA, 61.525μA].

If forward diode current = 6mA the determine the applied voltage across the diode?[0.669V]

Assume Ideality factor = 1 .

Question 5. If reverse saturation current = 5fA then determine I_D for V_D = - 0.04V and V_D =

-2V[Answer -3.92fA, -5fA].

Question 6. When I_D = 1.14mA what is V_D at 300K if I_{DO} = 5fA.[Ans.0.68V]

What is the diode voltage at275K and at 350K if Diode Current is to be maintained constant. [0.742V,0.555V]

Diode Forward Voltage V_D changes by -2.5mV per Kelvin for maintaining the forward diode current constant.

Question 7. Given η = 1,I_{DO} = 10fA and Thermal Voltage(V_{Th}) = 25mV. Determine diode current at Diode Voltage = 0.7V,0.65V,0.6V,0.4V,-0.1V and -5V.

[Answer:14.46mA,1.95mA,0.26mA,88nA,-9.8fA and -10fA]

Question8. Given $\eta=1$,$I_{DO}=100$fA and Thermal Voltage(V_{Th}) $=25$mV. Determine diode current at Diode Voltage$=0.4$V and determine diode voltage for diode current$=1$mA.

[0.888μA,0.576V]

Question 9. Given $\eta=1$,determine the factor by which current increases if forward diode voltage is increased by 100mV.[Hint:60mV increase gives a decade increase; Answer: factor of 46.4]

Question 10. In a Si Diode 1N4135, $I_{DO}(300K)=10nA$. Determine $I_{DO}(316K)$. [30.3nA]

Hint: $I_{DO}(316K)= I_{DO}(300K)\times 2^{(T\text{-}300K)/10K} = 10nA\times 2^{(316K\text{-}300K)/10K}=10nA\times 2^{1.6}=30.3nA$.

Question 11. At 300K, forward bias of a diode $=$ 0.7V and forward diode current $=$ 7.42mA. To maintain the current constant at 7.42mA determine V_D* at 316K. [660mV]

Question 12. In 1N4135, at 300K the current flowing through the diode is $I_D = 10mA$

given $I_{DO}(300K)= 10nA$ and $\eta=2$. Using the Real Diode Equation, forward diode Voltage is determined to be $V_D =0.7185V$.

If the forward diode current is to be maintained constant at 10mA then determine V_D at 290K, 310K and at 320K.

Solution : Using the Real Diode Equation we get:

$$V_D = \eta \frac{kT}{q} ln\left(\frac{10mA}{10nA \times 2^{\frac{T-300}{10}}} + 1\right)$$

Using the above equation we obtain forward diode voltage at T =290K, 300K, 310K and 320K.

The answers are:

Temperature(K)	290	300	310	320
V_D(V)	0.725	0.714	0.701	0.6855

Question 13. Given I_{DO}(300K) = 10nA,η = 2. Determine I_D at V_D = 0.6V and at 0.7V. [Ans. 1.02mA and 7mA]

If I_D = 5mA then determine V_D at 300K. [Ans. 0.682V]

Question 14. In a given Si-Diode, I_{DO}(300K) = 1nA,η = 2. At 300K, determine I_D at V_D = 0.7V, 0.1V, 0V, -0.1V and at -0.7V.

[Ans. 0.702mA, 5.8nA, 0A, -0.854nA, -1nA]

Question 15. In a given Si-Diode, $I_{DO}(300K) = 1nA, \eta = 2$. At 300K, determine V_D

at I_D = -0.5nA, 0A, 0.5nA, 50nA and 50μA.

[Ans. -0.036V, 0V, 0.021V,0.204V, 0.563V]

Question 16.(a)Determine built-in barrier potential Φ_{BO} in a Si-diode. Silicon packing density is 5×10^{22}atoms/cc. P-Side doping is 1Acceptor Atom per 10^8 of Si Atoms and N-Side doping is 2 Donor Atoms per 10^7 of Si Atoms.

Hint: P-Side doping is 1Acceptor Atom per 10^8 of Si Atoms implies $N_A = (5 \times 10^{22}/10^8)/cc = 5 \times 10^{14}/cc$.

N-Side doping is 2 Donor Atoms per 10^7 of Si Atoms implies $N_D = (5 \times 10^{22}/(10^7/2))/cc = 1 \times 10^{16}/cc$.

[Ans.0.64V]

(b) Determine built-in barrier potential Φ_{BO} in a Si-diode where :

	P-side	N-side
σ	50S/m	100S/m
μ	0.05m2/(V-s)	0.13m2/(V-s)

Hint:

$$\sigma_p = q\mu_p p_p^- = q\mu_p N_A \qquad therefore\ N_A = 6.24 \times 10^{21}/m^3$$

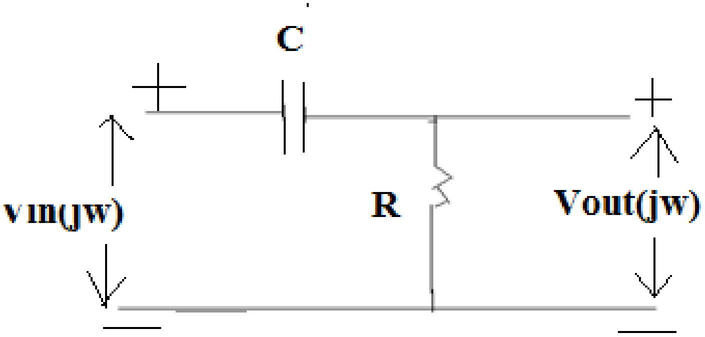
$$\sigma_n = q\mu_n n_n^- = q\mu_n N_D \qquad therefore\ N_D = 4.8 \times 10^{21}/m^3$$

Therefore $\Phi_{BO} = 0.687$V.

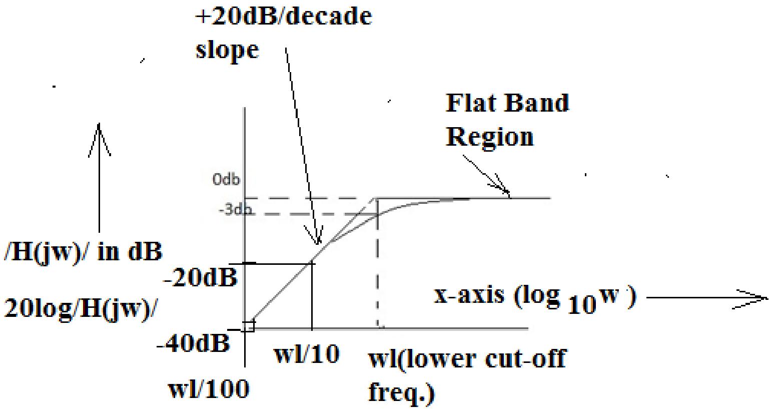
SSPD_Experiment 2_Figures
This module gives the Figures of SSPD_Experiment 2.

Experiement -2-Figures of HPF.

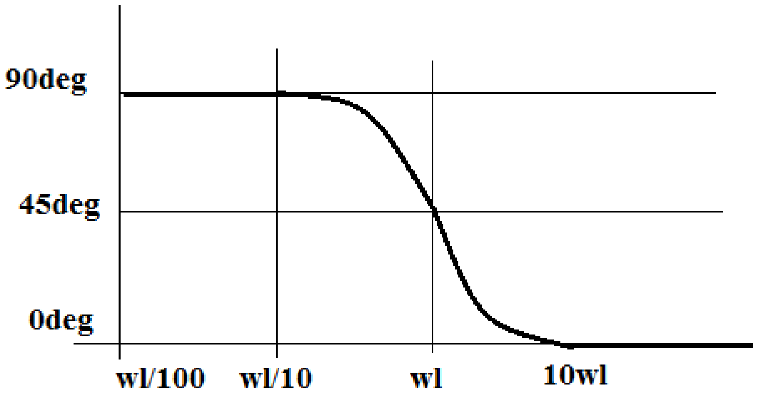
2.4. High pass RC circuit (as shown in Figure 1.a.)



(a) R-C configuration of HPF



(b) Bode Plot of the Transfer Function Magnitude /H (jw)/ with respect to the log to the base 10 of Frequency.



(c) Phase vs Frequency Response of HPF

Figure 1.(a) Circuit Topology of High Pass Filter.

(b) Magnitude of the transfer function in dB(decibels) vs Frequency(logarithmic scale) of High Pass Filter(HPF)

(c) Phase of the transfer function in dB(decibels) vs Frequency(logarithmic scale)of HPF.

The two plots b and c together are known as Bode Plot.

SSPD_Chapter 4_Bipolar Junction Transistor Physics
This Section gives the insight into the Forward Active Mode, Inverse Active Mode , Saturation Mode and Cut-Off Mode. of Bipolar Junction Transistor.

4. Bipolar Junction Transistor.

The first 3 terminal Solid State Device to be invented and commercialized was Bipolar Junction Transistor. It was invented in 1948 and discrete Ge-BJT wre immediately mass produced. Subsequently the merits of Si-BJT were recognized and Si-BJT went into mass production for various Signal Processing Applications.In 1961 first Bipolar Si-IC also went into mass production. This was Resistance-Transistor-Logic IC. MOSFET IC's were commercialized in 1970 mainly for application in watches but subsequently they became the main stay of Micro-processors chips. NMOS was soon replaced by CMOS because of its nano-watt standby dissipation characteristics. Today the heat management problems at nano-scale has led to the adoption of CMOS ICs but BJT still finds application in very high frequency,low noise and high output power applications.It is used in mobile phone output stages. BJT Analog ICs will continue to be the interface circuits in most general applications.

Bipolar Junction Transistor (BJT) is a back-to-back(NPN) or face-to-face device(PNP) and it is three terminal device as shown in Figure 4.1.

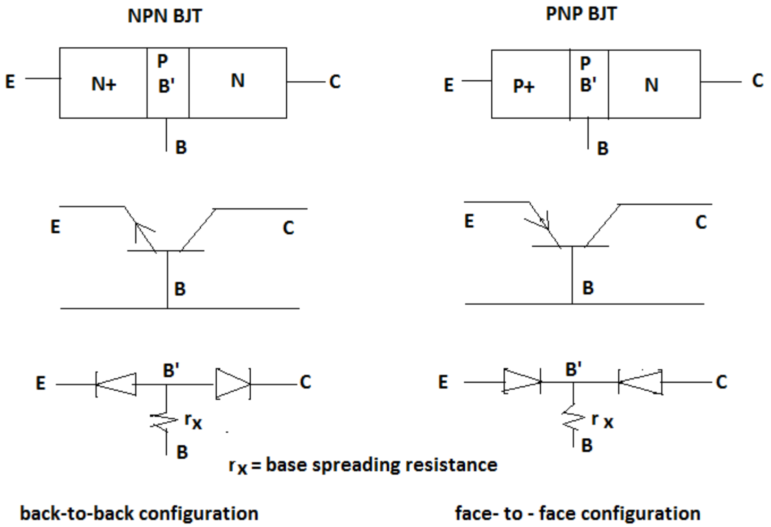

NPN BJT
PNP BJT
E
N+
P
B'
N
C
B
P+
r_x
r_x = base spreading resistance
back-to-back configuration
face- to - face configuration

Figure 4.1. Basic Structure, Symbol and D.C. Model of BJT.

It has four modes of operation as given below:

Forward Active Mode: EB Jn Forward biased, BC Jn Reverse biased.

Inverse Active Mode: EB Jn Reverse biased, BC Jn Forward biased.

Saturation Mode: EB Jn Forward biased, BC Jn Forward biased.

Cut off Mode: EB Jn Reverse biased, BC Jn Reverse biased.

It can be biased as a linear device to amplify signals. It can be biased as a switch to act as Digital Logic

Gate.

In Forward Mode: BJT acts like a linear device.

In Saturation Mode(short circuit) and in Cut-off Mode(open circuit) it acts like a Switch.

Figure 4.2. gives Carrier Concentration Profile in Forward Active Mode.

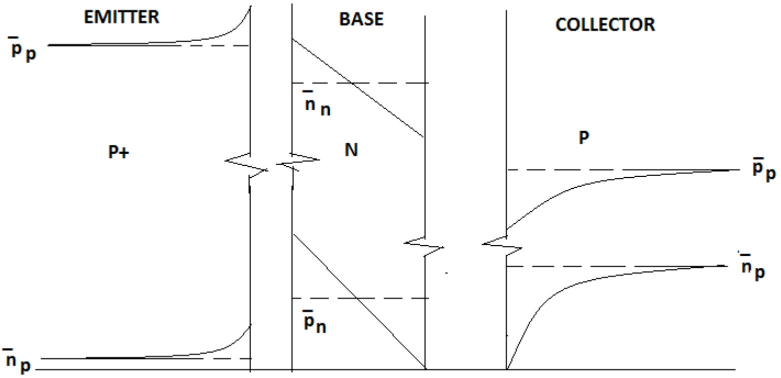


Figure 4.2. Carrier Concentration Profile in Forward Active Mode in PNP BJT

Figure 4.3. shows the carrier flow. In Figure 4.3. we note that barrier potential at Forward Biased EB Junction reduces from Φ_{BO1} to Φ_{BO1}- □V_{EB}□ and at Reverse Biased BC junction it increases from Φ_{BO2} to Φ_{BO2}- □V_{CB}□. Therefore holes are easily injected into the Base. The Base is very narrow, less than 1μm, hence the injected holes easily diffuse across the BASE and fall down the reverse biased BC potential gradient. Hole is minority carrier in N-Type Base hence BC Junction acts as down-hill gradient for the holes. In effect Emitter is injecting the holes into the Base and Collector is collecting the holes and holes are transiting the BASE in a finite time called Base Transit Time = τ. This Transit Time decides the Transit Frequency or the Unity Gain Frequency of the given BJT. This Transit angular Frequency $\omega_T = (1/\tau) = 2\pi f_T$. This Transit Frequency ranges from 200MHz to 800MHz for UHF BJT and it can go as high as 500GHz for Microwave Transistors.

$$\alpha_F = Forward\ Current\ Transfer\ Ratio = \frac{I_{Ep}}{I_E} \times \frac{I_{Cp}}{I_{Ep}} \times \frac{I_C}{I_{Cp}} = \frac{I_C}{I_E} = \gamma . \beta^* M \qquad 4.1.1$$

Here γ = Injection Efficiency = I_{Ep}/I_E ;

And β^* = Base Transport Factor = I_{Cp}/I_{Ep} ;

M = Avalanche Multiplication Factor.

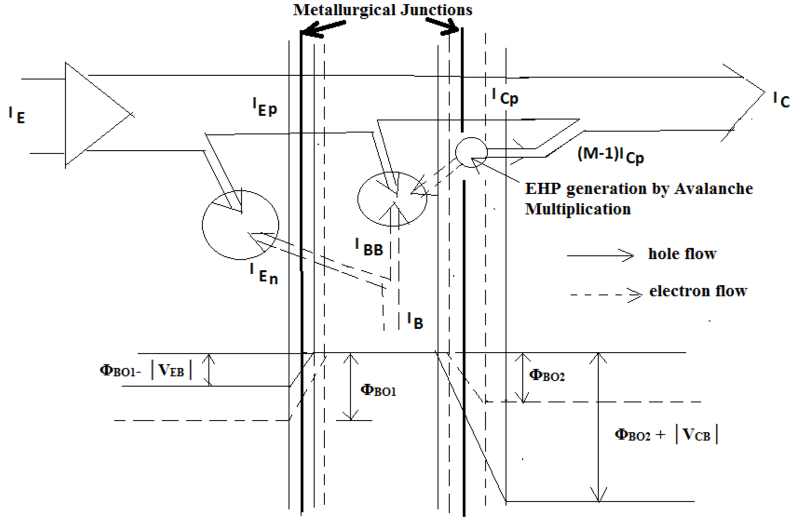


Figure 4.3. Carrier Flow in a Forward Active Biased BJT.

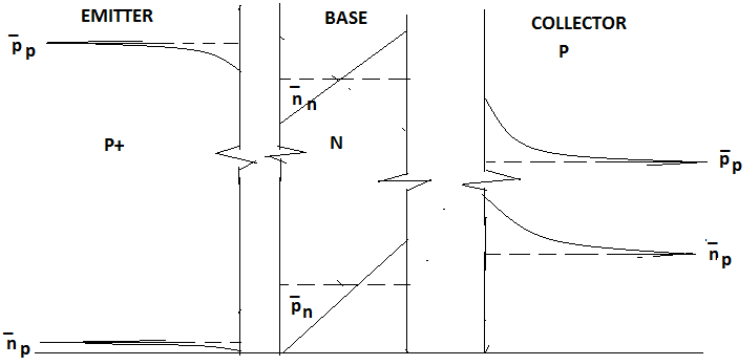


Figure 4.4. Carrier Concentration Profile of Inverse-Active Mode.

In Figure 4.4. Carrier Concentration Profile of Inverse Active Mode is shown. It is exactly the inverse of Forward Active Mode. The performance of Inverse Active Mode BJT is much inferior to that of Forward Active Mode as we will see this in the Output Family of Curve of Inverse Active Mode

Common Base BJT.

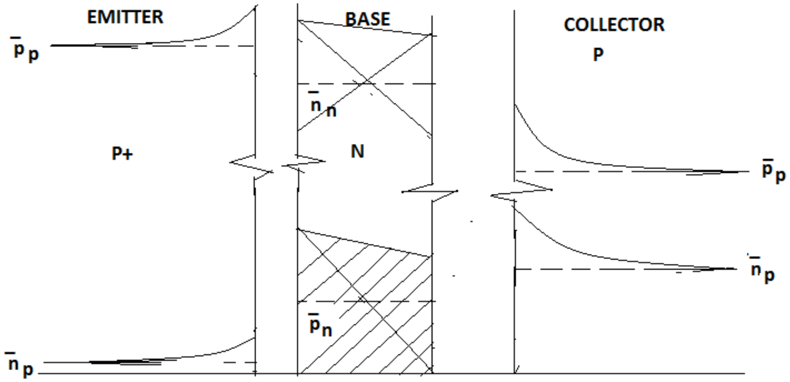


Figure 4.5. Carrier Concentration Profile in Saturation Mode.

In Figure 4.5. Carrier Concentration Profile in Saturation Mode is shown. Since both junctions are forward biased hence holes are being injected from Emitter and Collector both. This leads to heavy build up and storage of minority carries. It causes a very large Storage Time Delay during switching. Hence we use Schottky Clamped BJT where deep saturation is prevented. This leads to drastic improvement in switching speed. In 70's gold dope was used to kill the minority carruier and thereby reduce the storage delay. But it led to large reverse leakage current. Hence Schottky Clamped method was adopted for improving the switching speed.

In Figure 4.6.Carrier Concentration Profile in Cut-off Mode is shown. Both Junctions are reverse biased hence no current flows..

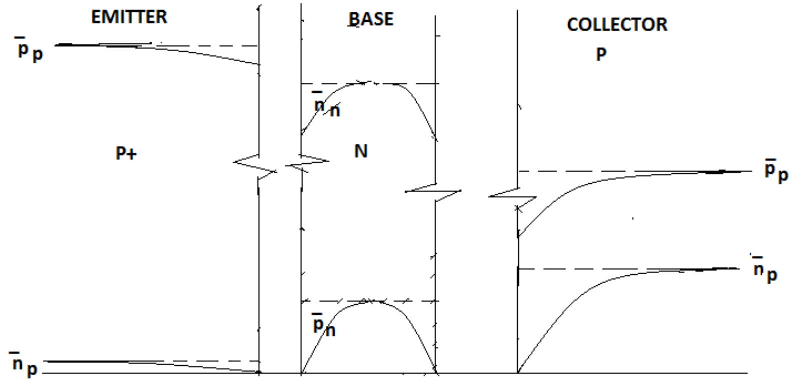


Figure 4.6. Carrier Concentration Profile in Cut-off Mode.

Analog Electronics Lecture 2_PartB_BJT configurations & modes

Abstract: This describes three circuit configurations, four modes of operation and five schemes of biasing;

Analog Electronics Lecture 2_PartB_BJT configurations & modes

Key words;CE,CC & CB BJT;

Abstract: This describes three circuit configurations, four modes of operation and five schemes of biasing;

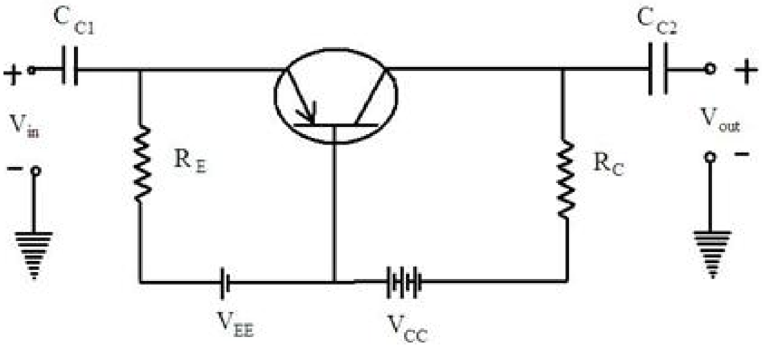
BIPOLAR JUNCTION TRANSISTOR

FOUR MODES OF OPERATION

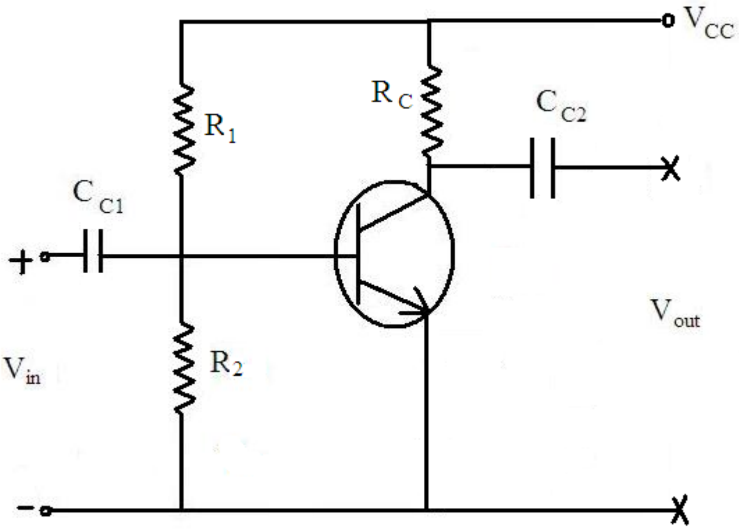
Type Of Biasing	Emitter-Base Junction	Base-Collector Junction
Forward Active Mode	Forward Biased	Reverse Biased
Inverse Active Mode	Reverse Biased	Forward Biased
Saturation Mode	Forward Biased	Forward Biased
Cut-Off Mode	Reverse Biased	Reverse Biased

THREE CIRCUIT CONFIGURATIONS

1. COMMON BASE CONFIGURATION



1. COMMON EMITTER CONFIGURATION



1. COMMON COLLECTOR CONFIGURATION

The Figure of Common Collector Configuration is not given here. An emitter follower is an example of Common Collector Configuration. In Emitter Follower the collector is directly connected to the positive of the battery and under incremental analysis or under small signal analysis the battery is considered short circuited.Hence under signal condition, collector is effectively shorted to ground. Hence Emitter Follower is a Common Collector Configuration.In Emitter followr , input is at base but output is at emitter.The voltage at emitter follower follows the voltage at the input namely the base. Hence the voltage gain is Unity but current gain is almost 100. It can deliver a large current. Hence it is a Buffer and used for driving the load after the pre-amplifier.

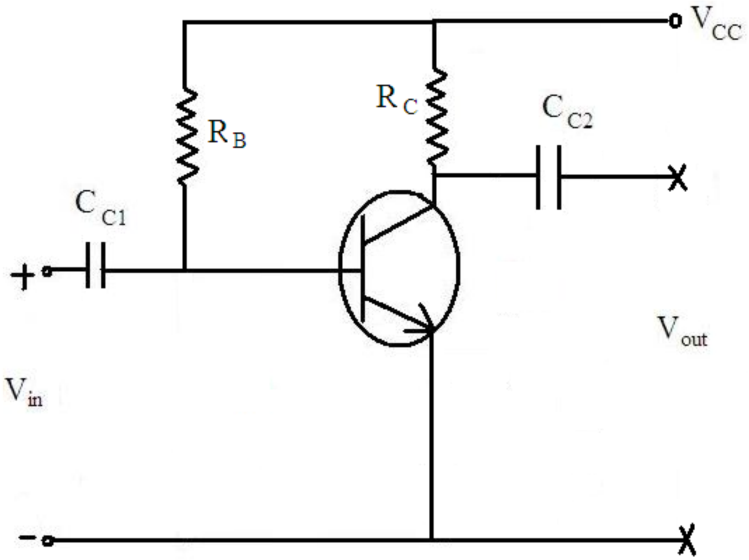
TYPES OF BIASING SYSTEM USED

(1)Two battery biasing:

The 3 circuit configurations shown in the previous page are the examples of this biasing.

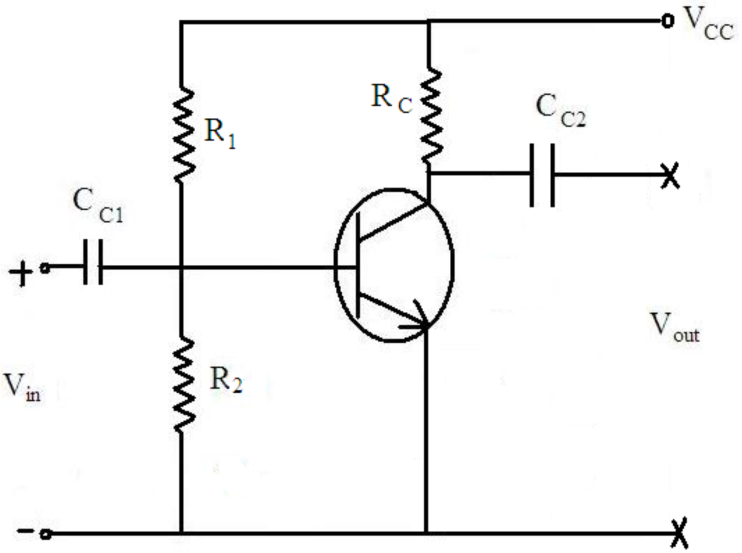
(2)Fixed Bias:

This type of biasing system offers poor stability because it has negative feedback..



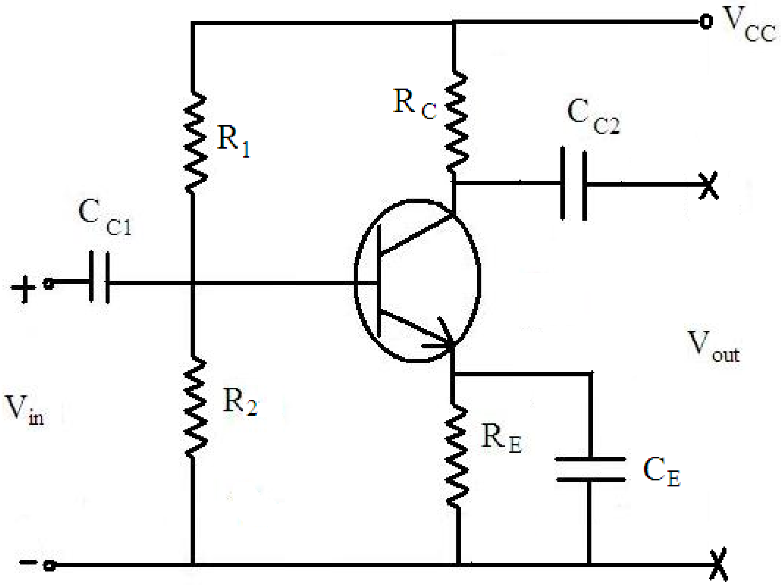
(3)Potential Divider Bias:

This type of biasing also offers poor stability because this also has no negative feedback.



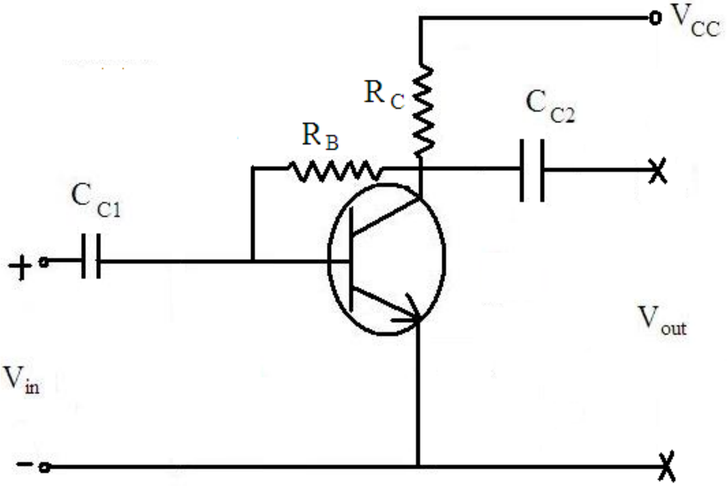
(4)Self Biasing:

This type of biasing system offers high stability because of current series feed-back.



(5)Widlar Biasing:

The above types of biasing systems are used normally for discrete circuits. But this type of biasing is used for IC technologies. It offers high stability because of voltage shunt feed-back.This is also known as collector to base feed-back.


V_CC
R_C
C_C2
R_B
C_C1
+
V_out
V_in
-

Analog Electronics Lecture 2_PartC_I-V output characteristics of BJT

Abstract: This describes the D.C. parameters of BJT.

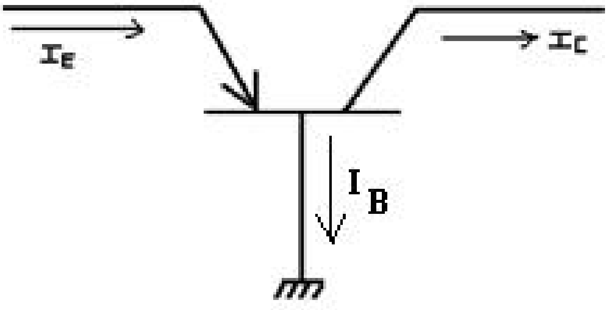
Analog Electronics Lecture 2_PartC_I-V output characteristics of BJT

Key words;BJT;

Abstract: This describes the D.C. parameters of BJT.

BJT Common Base Configuration:

For CB Configuration :



$I_C = \alpha_F M I_E + I_{CBO}$

Where α_F = D.C. Forward current transfer Ratio of CB BJT = I_C/I_E ;

M = Avalanche Multiplication Factor at Base-Collector Junction given by = 1/[1-(Vcb/BVcbo)^n]

$$BV_{CBO}$$

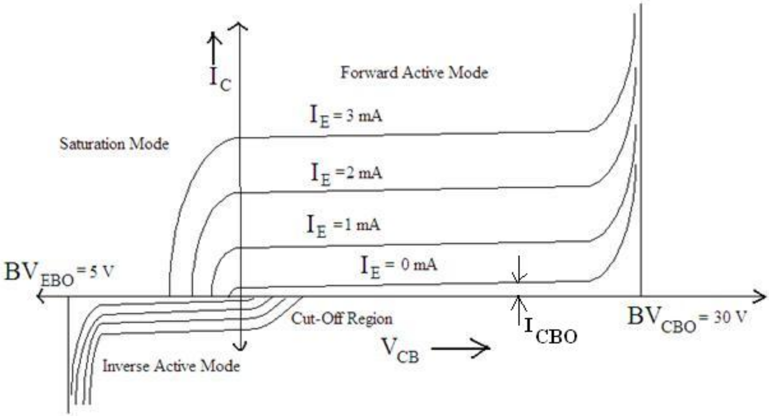
Where n = Miller Indices = 2~6

I_{CBO} = Reverse leakage current at CB Jn with emitter open.

$$BV_{CBO}$$

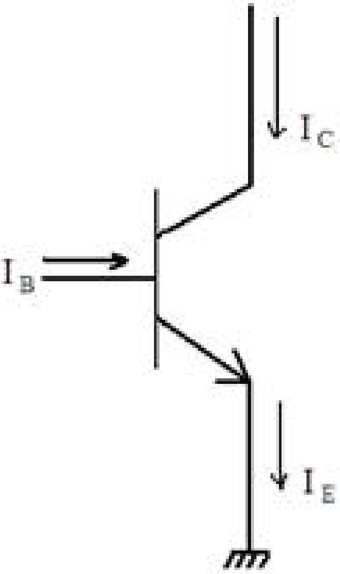
= Avalanche breakdown voltage at CB Jn with emitter open

OUTPUT CHARACTERISTICS OF Common Base Bipolar Junction transistor



$I_C = I_{CBO}$ when $I_E = 0$ mA. This is the reverse leakage current at CB Junction with Emitter open and is of nA range.

BJT Common Emitter(CE) Configuration:



For CE Configuration :

$I_C = \alpha_F M I_E + I_{CBO}$

$=> I_C = \alpha_F M(I_B + I_C) + I_{CBO}$

$=> I_C(1-\alpha_F M) = \alpha_F M I_B + I_{CBO}$

$$=>I_C = \frac{\alpha_F M\ I_B}{(1-\alpha_F M)} + \frac{I_{CBO}}{(1-\alpha_F M)} \qquad \textbf{Eq.2}$$

$$\boxed{=>I_C = \frac{\alpha_F M\ I_B}{(1-\alpha_F M)} + I_{CEO}} \qquad \textbf{Eq.3.}$$

$$\text{Where } I_{CEO} = \frac{I_{CBO}}{(1-\alpha_F M)}$$

If $M=1$,

$$\frac{\alpha_F M}{(1-\alpha_F M)} = \frac{\alpha_F}{1-\alpha_F} = 100$$

Where $\alpha_F = 0.99$.

At low voltages we have M=1.

$$=>I_C= \frac{\alpha_F}{(1-\alpha_F)} I_B + \frac{I_{CBO}}{(1-\alpha_F)} \qquad \textbf{Eq.4.}$$

If we define $\frac{\alpha_F}{(1-\alpha_F)} = \beta_F$ =D.C. Short Circuit Current gain of CE BJT

$$= \frac{I_C}{I_B}$$

Then we get:

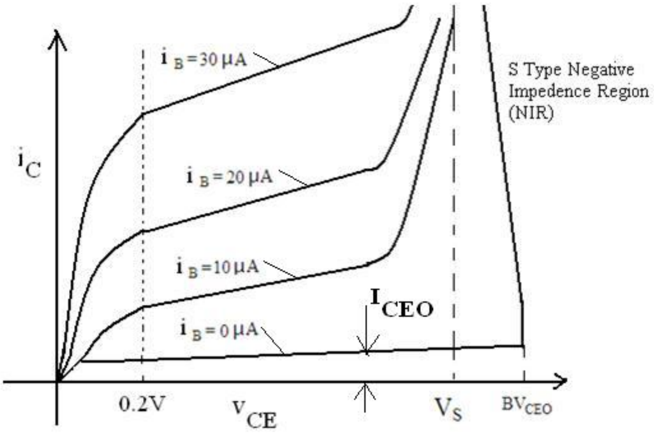
$$=>I_C= \beta_F I_B + (1+\beta_F) I_{CBO} \qquad \textbf{Eq.5}$$

That is:

$$=>I_C= \beta_F I_B + I_{CEO} \qquad \textbf{Eq.6}$$

Where $I_{CEO} = (1+\beta_F) I_{CBO}$

OUTPUT CHARACTERISTICS OF Common Emitter Bipolar Junction transistor



NOTE:-The slope in the figure is due to base width modulation which is also known as Early Effect.

$I_C = I_{CEO}$ when $I_B = 0$ mA. This is the collector junction leakage current at CB Junction with Base open and is of μA range.

Let us consider :

$$I_C = \frac{\alpha_F M\, I_B}{(1-\alpha_F M)} + \frac{I_{CBO}}{(1-\alpha_F M)} \qquad \textbf{Eq.7}$$

If $\alpha_F M = 1$, then

$$\frac{I_{CBO}}{(1-\alpha_F M)} = \infty$$

At this point , break over occurs. And we have

BV_{CEO} = Break-over Voltage with Base Circuit open.

When $\alpha_F M = 1$

That is

$$\alpha_F \left[\frac{1}{1 - \left(\frac{V_{CB}^*}{BV_{CBO}} \right)^n} \right] = 1 \qquad \textbf{Eq.8}$$

$$\alpha_F = 1 - \left(\frac{V_{CB}^*}{BV_{CBO}} \right)^n$$

$$\left(\frac{V_{CB}^*}{BV_{CBO}} \right)^n = 1 - \alpha_F$$

But :

$$1 - \alpha_F = \frac{1}{1 + \beta_F} \qquad \textbf{Eq.9.}$$

Thus :

$$\left(\frac{V_{CB}^{*}}{BV_{CBO}}\right) = \frac{1}{\sqrt[n]{1+\beta_F}}$$ **Eq.10**

Thus:

$$V_{CB}^{*} = \frac{BV_{CBO}}{\sqrt[n]{1+\beta_F}}$$ **Eq.11**

Putting the required Values i.e. $BV_{CBO} = 30V$, $\beta_F = 100$, we get $V_{CB^*} = 18$ V = BVceo ;

Now we have to know more about BV_{CES} (Breakover Voltage when the base circuit is shorted)

$BV_{CBO} > BV_{CEX} > BV_{CES} > BV_{CEO}$

Where $BV_{CBO} =$ Breakover Voltage of the collector base junction when the emitter circuit is open.

$BV_{CEX} =$ Breakover Voltage of CE BJT for a given termination R_X at the base

$BV_{CEO} =$ Breakover Voltage of CE BJT when the base circuit is open.

$BV_{CES} =$ Breakover Voltage of CE BJT when the base circuit is shorted to ground.

$BV_{CEX} =$ Breakover Voltage when the base circuit is connected to ground through a source Resistance

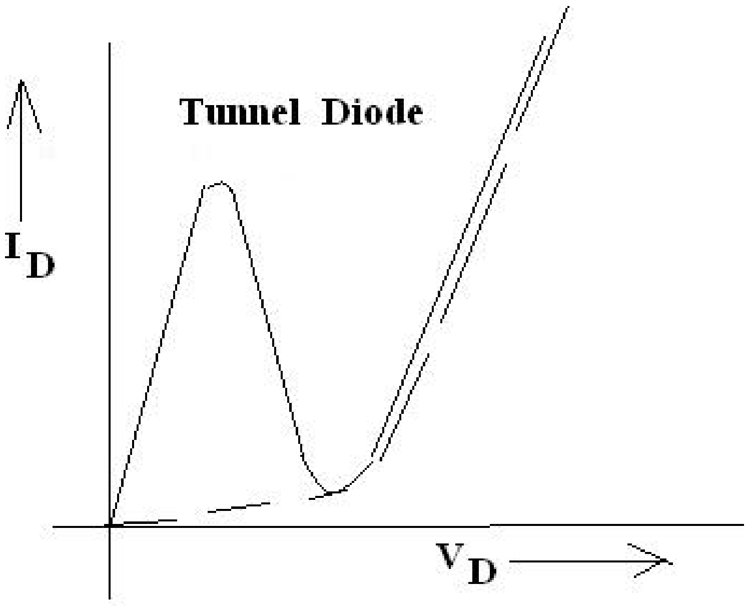
(R_S) .

By proper base termination, the permissible region of operation can be extended upto BV_{CBO}.

Thus we have seen that breakover occurs at $\alpha_F M = 1$. At low current α_F is very small, almost about 0.1.Therefore voltage has to be taken to a large value to satisfy $\alpha_F M = 1$. But as soon as breakover occurs large current starts flowing. With large current α_F improves from 0.1 to 0.99. Hence $\alpha_F M = 1$ is satisfied at lower voltage Vs. Therefore breakover curves settles down at Vs. This voltage V_S is known as sustaining voltage. Because of the fact that :

$V_S < BV_{CEO}$

We get a S Type Negative Impedance Region(NIR). In SCR and UJT also we get S Type NIR but in Tunnel Diode as shown in the Figure below we get N Type NIR.



COMPARISON BETWEEN COMMON BASE AND COMMON EMITTER CONFIGURATIONS

S.No.	COMMON BASE	COMMON EMITTER
1	h_{rb}(reverse transmission factor)~10^{-5}Thus it behaves as a near unilateral device.	h_{re}~10^{-4}Thus it behaves as a non-unilateral device.
2	In RF	At low

	applications the circuit has a high probability of parasitic oscillations but in CB because of near unilaterality , probability of parasitic oscillation goes down. Hence for RF applications CB is the preferred circuit configuration.	frequencies there is no danger of parasitic oscillations hence CE can be used even with poor reverse transmission factor.
3	h_{ob} = 1/(2M)Thus it behaves as a near ideal current source. Thus it is very suitable for charging a capacitance with a constant current to generate a saw-tooth waveform.	h_{oe} = 1/(40K)Thus it behaves as a non-ideal current source.

Both CB and CE are Current Controlled Current Source. CB is a near ideal CCCS whereas CE is a non-ideal CCCS.

AnalogElectronics_Lecture4_PartA_dc,low&high frequency model of CB BJT.
Lecture 4 deals with low frequency and high frequency model of CB and CE BJT. In part A, I present the EberMoll Model. From Eber Moll Model I derive the T-Model of CB BJT at low as well as high frequency model.

AnalogElectronics_Lecture4_PartA_dc,low&high frequency model of CB BJT.

The DC model of BJT.

PNP is a face to face diode.

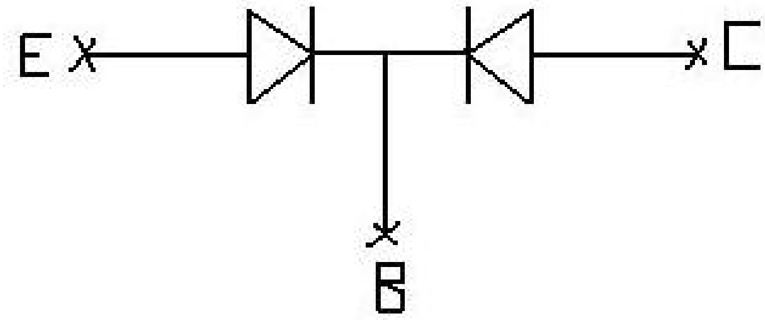


Figure 1.PNP is a face to face diode configuration.

NPN is a back to back diode.

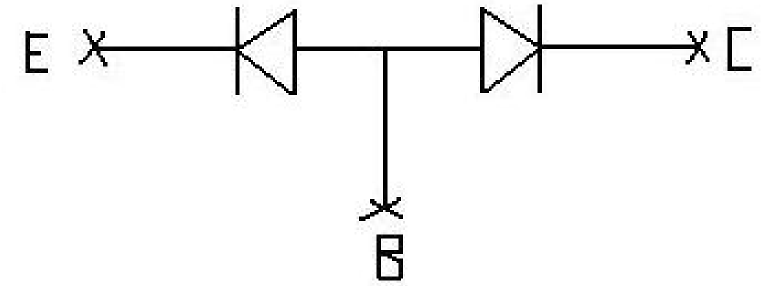


Figure 2. NPN is a back to back diode

DC Biasing of CB BJT (Common Base BJT) In the figure base is grounded.

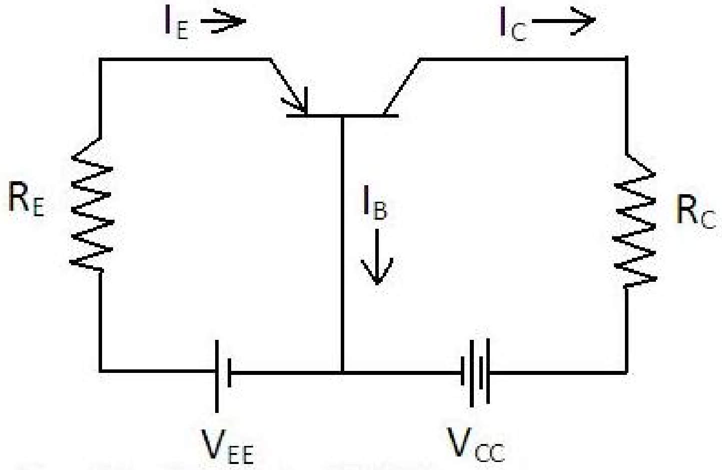


Figure 3. Two Battery biasing of CB BJT.

The DC model of CB BJT is called Eber Moll Model. It is a large signal model.

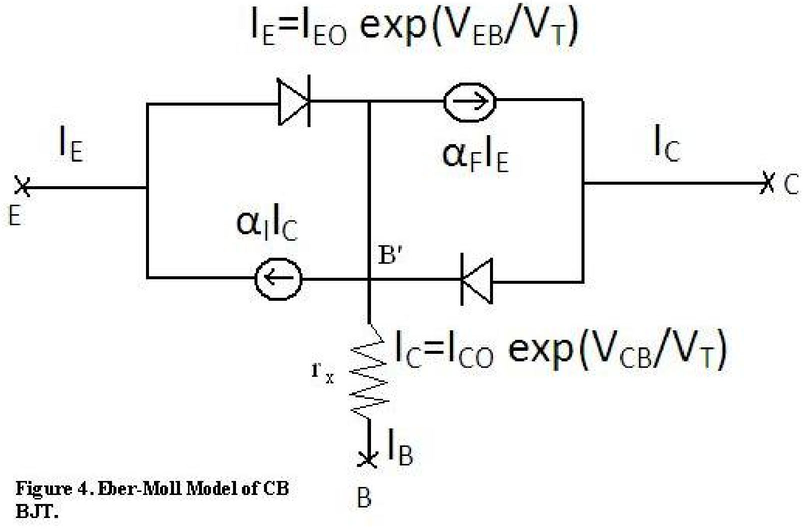

$I_E = I_{EO} \exp(V_{EB}/V_T)$
I_E
E
$\alpha_F I_E$
I_C
C
$\alpha_I I_C$
B'
$I_C = I_{CO} \exp(V_{CB}/V_T)$
r_x
I_B
B

Figure 4. Eber-Moll Model of CB BJT.

Figure 4. Eber Moll Model of PNP Transistor under no bias.

As can be seen from the Eber Moll Model:

$\alpha_F I_E$ is the controlled current source controlled by the current injected from E to B where α_F is short circuit forward current transfer ratio.

$\alpha_I I_C$ is the controlled current source controlled by the current injected from C to B where α_I is short circuit inverse current transfer ratio.

Because of asymmetrical doping density, [Heavy Emitter doping(10^{19}/cc), Intermediate Base doping(10^{17}/cc) and light Collector doping(10^{16}/cc)] emitter to base injection efficiency is 99%

whereas collector to base injection efficiency is less than 10% therefore ($\alpha_I = 0.1) << (\alpha_F = 0.99$).

Under Forward Active Mode:

EB diode is forward biased and CB diode is reversed biased. Hence current is being injected from emitter to base but no current is being injected from collector to base. Therefore EB diode is shown and CB diode has been omitted but there is reverse saturation current/reverse leakage current I_{CBO} flowing through reverse biased CB junction. In fact there are two currents flowing across the CB junction. One is the leakage current I_{CBO} and the other is the useful transistor current $\alpha_F I_E$ due to transistor action at EB junction.

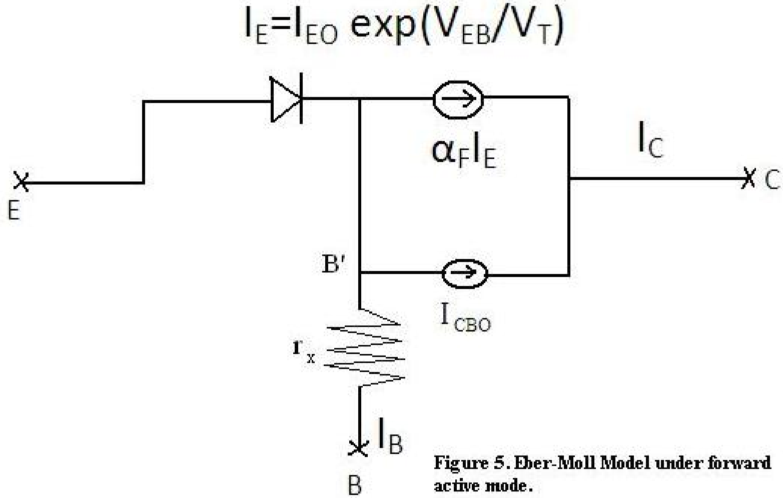


Figure 5. Eber-Moll Model under forward active mode.

Figure 5. Eber Moll Model under Forward Active Mode.

LOW FREQUENCY MODEL OF CB BJT.

A capacitively coupled CB Amplifier is shown in Figure 6. It amplifies sinusoidal signal but blocks dc signal.

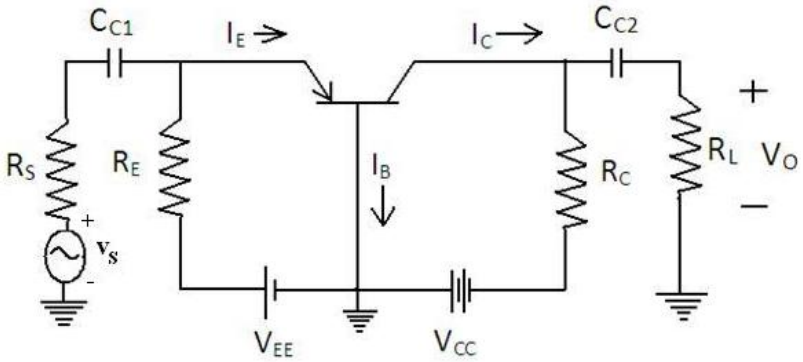


Figure 6. RC-coupled CB BJT Amplifier with two battery biasing.

In Figure 6 we have capacitively coupled source and capacitively coupled load in CB BJT.

In Figure 6 we show a CB BJT Amplifier. Because of near unilaterality of CB BJT (h_{rb} negligible), it is very suitable for RF tuned amplifiers. RF tuned amplifiers have a serious problem of parasitic oscillation because of output - input interaction and this interaction is due to h_r(reverse transmission factor).

CE BJT has $h_{re} = 10^{-4}$ hence CE BJT Amplifier is very prone to parasitic oscillations. But CB BJT has $h_{rb} = 10^{-5}$ hence CB BJT is the configuration of choice for implementing RF tuned amplifiers.

Also capacitive coupling is used because once a Quiescent Point or DC operating point is set up we do not want it to be disturbed by load or source. That is load and source must be a.c. coupled but d.c. decoupled. Capacitive coupling serves this end.

Capacitively coupled Amplifiers (CE, CB & CC) are always ac amplifiers. But inverting amplifiers made of Op Amp with dual power supply is a direct coupled amplifier. There are no coupling capacitors. Hence they are dc amplifiers. DC amplifiers can amplify from dc to high frequencies.

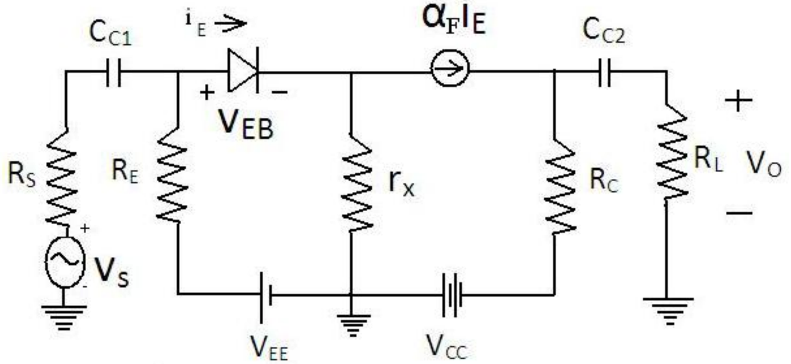


Figure 7. DC Model of CB BJT.

In Figure 7 we have the instantaneous model of CB Amplifier with signal coupled to it.

Therefore i_E (instantaneous emitter current) = I_E(emitter quiescent current) + i_e(emitter incremental current);

Similarly we have $i_C = I_C + i_c$ and $i_B = I_B + i_b$;

Also $I_C = \beta_F \times I_B$ and $i_c = \beta_f \times i_b$;

Here β_F = dc short circuit current gain and β_f = incremental short circuit current gain.

In BJT : $\beta_F = \beta_f$;

The instantaneous model also contains $r_x = r_{bb'}$ = base spreading resistance.

This is an ohmic resistance offered by the thin layer of P type silicon sand-witched between Emitter layer and Collector layer as shown in the cross sectional view of IC vertical npn BJT (Figure 8)

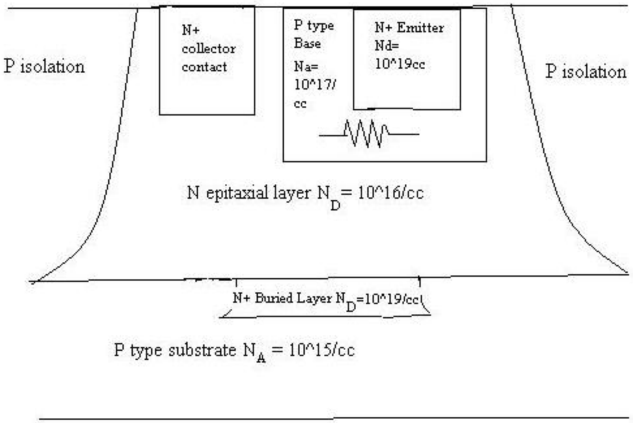

P isolation
N+ collector contact
P type Base
Na= 10^17/ cc
N+ Emitter Nd= 10^19cc
P isolation
N epitaxial layer N_D = 10^16/cc
N+ Buried Layer N_D=10^19/cc
P type substrate N_A = 10^15/cc

Figure 8 Cross-sectional view of Vertical NPN IC BJT.

Figure 8. Cross sectional View of Integrated Circuit Vertical NPN BJT.

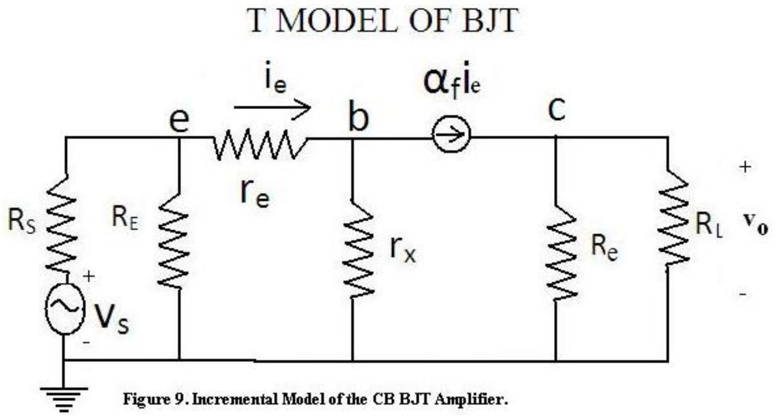


Figure 9. Incremental Model of CB BJT Amplifier. (the controlled current source is $\alpha_f i_e$).

This incremental model is known as the T-model of CB BJT.

As in incremental diode circuit, here also we short the DC Voltage Sources(V_{CC} and V_{EE}), we open circuit DC Current sources such as ($I_{CBO} + \alpha_F I_E$) but we retain the incremental current source $\alpha_f i_e$.

The coupling capacitances are shorted.

EB Diode is replaced by incremental resistance $r_e = (V_T/I_{EQ})$.

The ohmic resistances are retained as they offer resistance to DC as well as incremental component of the current.

HIGH FREQUENCY MODEL OF COMMON BASE BJT

Under high frequency condition the Junction Capacitances associated with the depletion layers EB and BC and Diffusion Capacitance associated with minority carriers stored in base region start effecting the performance of the BJT. These parasitic capacitances effect the Frequency Response as well as the short circuit current gain.

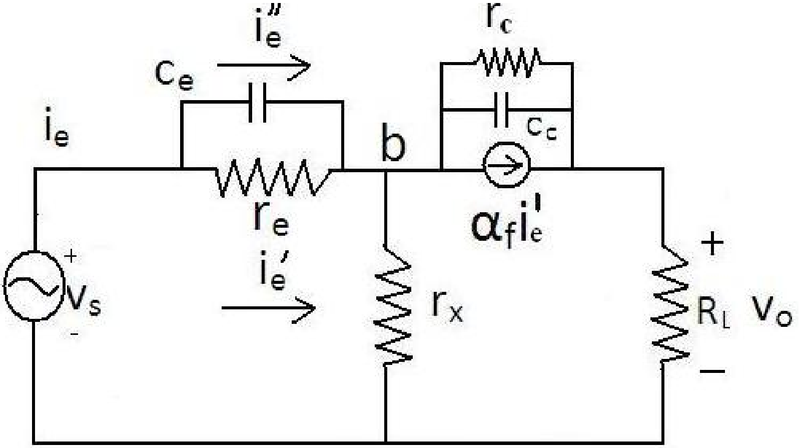


Figure 10. High Frequency T-model of CB BJT.

Figure 10. Incremental T-Model of the CB Amplifier Circuit at high frequency. (the controlled current source is αfie)

At high frequency diffusion and junction capacitances come into play.

$$C_e = C_{jEB} + C_D$$

$$C_c = C_{jBC} = C_{BO}$$

$=$ CBO This is generally 5pF.

$$(C_D)(r_e) = \tau_t$$

CD $=$ more than 70pF.

Where

$$\tau_t$$

is the transit time across the base.

$$C_{jEB} = \frac{\epsilon_o \epsilon_r A_1}{d_{EB}}$$

$$C_{jBC} = \frac{\epsilon_o \epsilon_r A_2}{d_{BC}}$$

$$d_{EB}$$

= Depletion layer width at EB Junction

$$d_{BC}$$

=Depletion layer width at BC Junction

$$A_1$$

=Cross sectional area of EB junction

$$A_2$$

=Cross sectional area of BC junction

$$i_e = i_e' + i_e''$$

Here

$$i_e'$$

is the useful transistor current which controls the output current sources (αfoie')and

$$i_e''$$

is the parasitic current.

$$i_e''$$

has no role to play in the transistor action

$$i_e''\ does\ not\ contribute\ to\ transistor\ action\ .$$

α_{fo} = short circuit incremental forward current transfer ratioat low frequency.

$$v_o = \left(\alpha_{fo} i_e' R_L\right)$$

This is true only when we consider r_c = infinity and C_C to be open.

This means this model is valid only till

$$\omega_T$$

/3. Because only till

$$\omega_T$$

/3 the above assumption is valid. Above

$$\omega_T$$

/3 this model gives incorrect result.

$$i'_e = \frac{i_e}{\left(r_e + \frac{1}{j\omega C_e}\right)} \times \frac{1}{j\omega C_e}$$

$$i'_e = \frac{i_e}{(1 + j\omega r_e C_e)}$$

$$v_o = \frac{\alpha_{fo} i_e}{(1 + j\omega r_e C_e)} R_L$$

$$i_c = \frac{\alpha_{fo} i_e}{1 + j\left(\frac{\omega}{\omega_\alpha}\right)}$$

$$\alpha_f = \frac{\alpha_{fo}}{1 + j\left(\frac{\omega}{\omega_\alpha}\right)}$$

Here α_f = short circuit current transfer ratio at all frequencies for CB BJT.

$$Here\ \omega_\alpha =$$

alpha cut-off frequency

$$\omega_\alpha = \frac{1}{C_e r_e}$$

$$\omega_\alpha \sim \frac{1}{\tau_t} = \omega_T$$

Here

$$\omega_T$$

is the transit frequency and

$$\tau_t$$

is the transit time.

The Bode Plot of the above expression of

$$\alpha_f$$

is given in Figure 12.

AnalogElectronics_Lecture4_PartB_low and high frequency model of CE BJT.
Part B of Lecture 4 derives the Hybrid-pi model of CE BJT from T-model of CB BJT at low and high frequencies.

AnalogElectronics_Lecture4_PartB_low and high frequency model of CE BJT.

INCREMENTAL MODEL OF CE BJT FROM T MODEL OF CB BJT

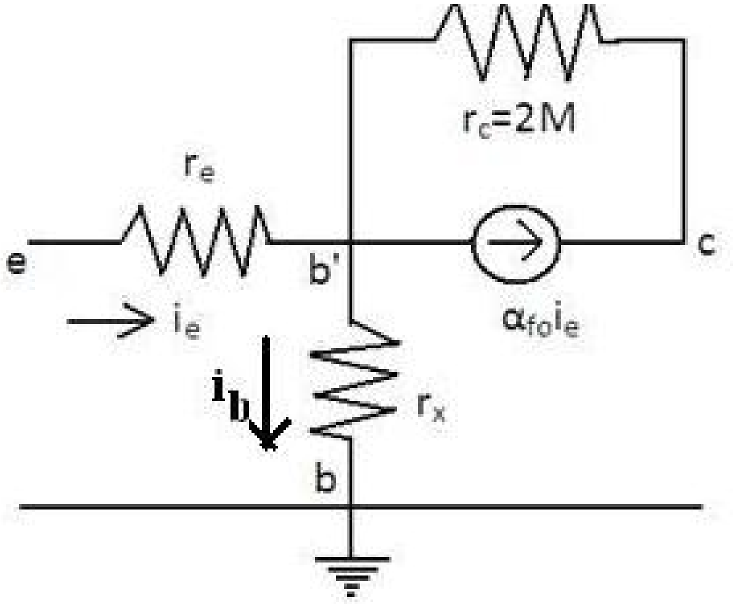


Figure 11a. The T-model of CB BJT.

Figure 11a. Low Frequency Incremental T Model of CB BJT.

Let us re-orient this as CE configuration.

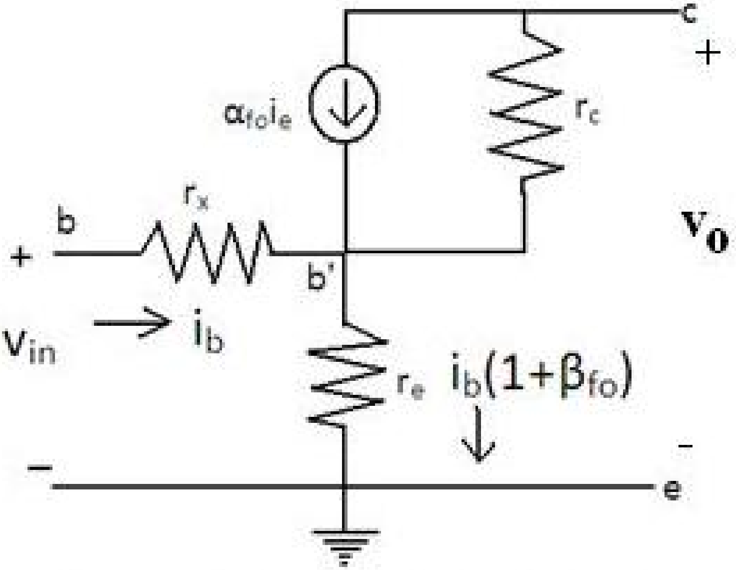


Figure 11b. Reoriented T-model of CB BJT.

Figure 11b. The reoriented T Model to represent CE BJT.

It should be noticed that in T-Model under normal orientation has controlled current($\alpha_f i_e$) coming out of collector node since base current is coming out of base node. But in reoriented T Model controlled current($\alpha_f i_e$) coming into collector node since base current is coming into base node.

Input Mesh Equation:

$$v_{in} = i_b \times r_x + (i_b(1 + \beta_{fo})r_e)$$

$$v_{in} = i_b[r_x + (1 + \beta_{fo})r_e]$$

$$(1 + \beta_{fo})r_e = r_\pi = \frac{\beta_o}{g_m}$$

Where g_m = trans conductance = I_C/V_T whereas r_e = V_T/ I_E and $I_C = \alpha_F I_E$

$$v_{in} = i_b[r_x + r_\pi]$$

The output current is =

$$i_C = \alpha_{fo} i_e = \beta_{fo} i_b$$

Here

$$\beta_{fo} = \frac{\alpha_{fo}}{1-\alpha_{fo}} \text{ or } \alpha_{fo} = \frac{\beta_{fo}}{1+\beta_{fo}}$$

.

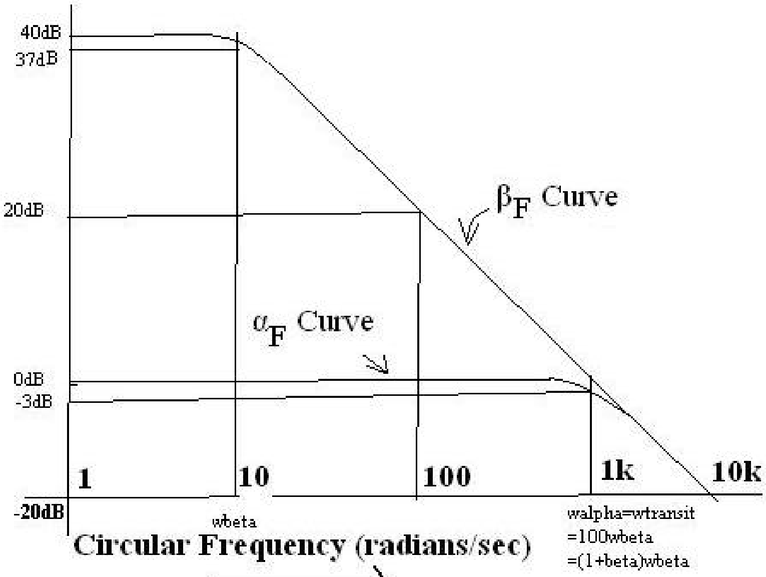


Figure 12. Bode Plot of alpha and beta.

Figure 12. Bode Plot of beta and alpha and location of ωβ and ωα.

Beta cutoff frequency = ωβ and alpha cutoff frequency = ωα .

Cut-off frequency = -3dB frequency

= this is the frequency where parameter falls to 0.707 of its flat band value or midband value

= corner frequency

= half power frequency

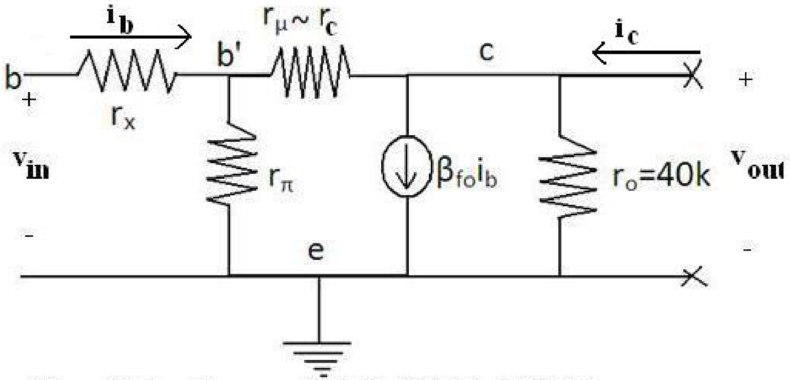


Figure 13. Low Frequency Hybrid-pi Model of CE BJT.

Figure 13. Low Frequency Hybrid-π Model of CE BJT.

$$r_o = \left(\frac{r_c}{1+\beta_o}\right)$$

This is due to EARLY EFFECT or due to Base Width Modulation. A parameter Early Voltage V_A is used for determining the output impedance of the hybrid –π Model. This output impedance is $1/h_{oe}$. The definition of Early Voltage V_A is given in Figure 14.

$$r_0 = \frac{1}{slope} = \frac{1}{\dfrac{I_{CQ}}{V_A + V_{CEQ}}}$$

$$r_o = \frac{V_A + V_{CEQ}}{I_{CQ}}$$

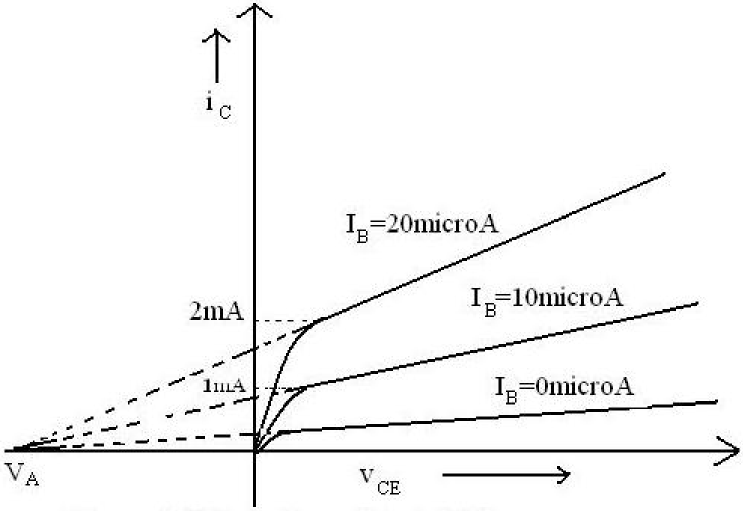


Figure 14 Illustration of Early Voltage.

Figure 14. The definition of Early Voltage for a CE BJT.

COMPARISON BETWEEN HYBRID-pi MODEL OF CE BJT AND T MODEL OF CB BJT

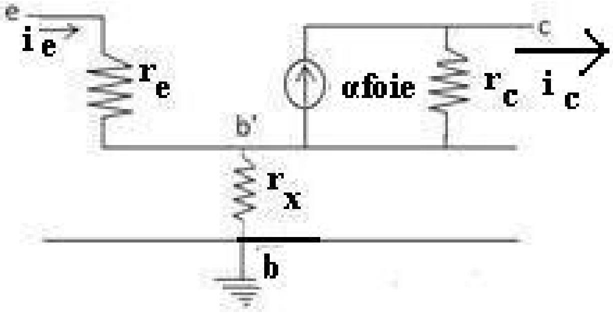


Fig 15a. T-Model of CB BJT.

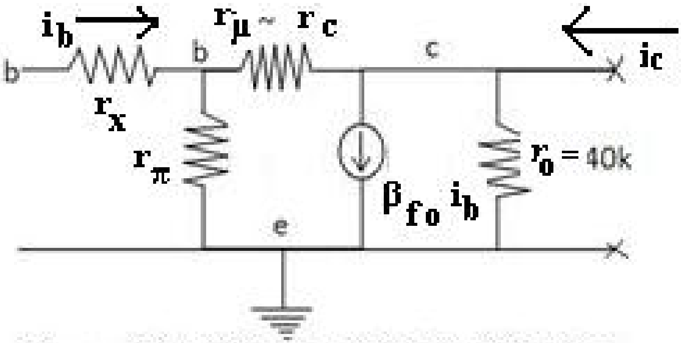


Figure 15b. Hybrid-pi Model of CE BJT.

T-model of CB BJT	Hybrid-pi Model of CE BJT
Unilateral model	Non Unilateral model
hrb = $\frac{r_x}{r_c} = \frac{100}{10M} = 10^{-5}$	

$$h_{re} = \frac{r_\pi}{r_\mu} = \frac{10^3}{10^7} = 10^{-4}$$

In CB BJT,if we consider base spreading resistance to be zero then

$$h_{ib} = r_e \quad h_{rb} = 0 \quad h_{fb} = -\alpha_{fo} \quad h_{ob} = \frac{1}{r_c}$$

$$\frac{v_o}{v_{in}} = \frac{\alpha_{fo} i_e R_L}{i_e r_e} = \alpha_{fo} \frac{R_L}{r_e} = A_{Vo}$$

$$R_{in} = r_e$$

$$R_{out} = r_c = \frac{1}{h_{ob}} = 10M$$

For CE BJT if we consider r_μ to be infinity then:

$$h_{ie} = r_x + r_\pi, h_{re} = 0, h_{fe} = +\beta_{fo},$$

$$h_{oe} = \frac{1}{r_o}$$

$$r_o = \frac{V_A + V_{CEQ}}{I_{CQ}}$$

$$v_o = -(\beta_{fo} i_b)(R_c \,||\, r_o)$$

$$\frac{v_o}{v_{in}} = \frac{(-\beta_{fo})(R_c \,||\, r_o)}{r_x + r_\pi}$$

$$A_{vo} = -g_m(R_c \,||\, r_o)$$

$$R_{in} = r_x + r_\pi$$

$$R_{out} = r_o$$

Incremental model at high frequency

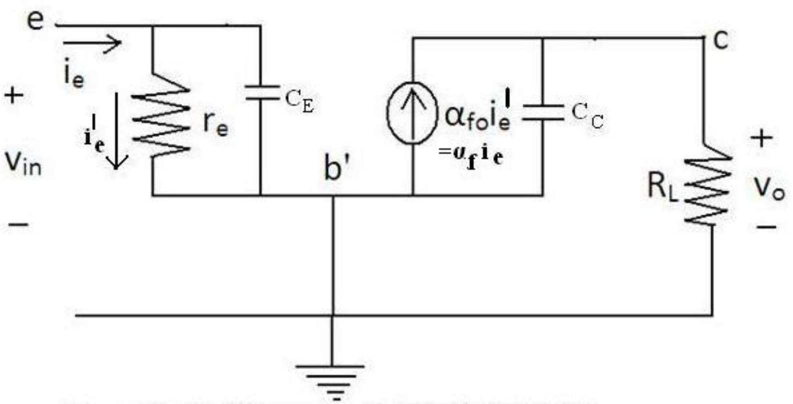


Figure 16. High Frequency T - Model of CB BJT.

Figure 16. High frequency T-Model of CB BJT.

$$\alpha_f = \frac{\alpha_{fo}}{1 + i\left(\frac{\omega}{\omega_\alpha}\right)}$$

$$\omega_\alpha \ is\ the\ \alpha\ cutoff\ frequency$$

$$\omega_\alpha \sim \omega_T = \frac{1}{\tau_t}$$

$$C_e = C_{jE} + C_D$$

$$\omega_\alpha = \frac{1}{r_e\left(C_{JE} + C_D\right)}$$

$$C_{jE} \ll C_D$$

$$\therefore \omega_\alpha = \frac{1}{r_e C_D}$$

$$\omega_\alpha = \frac{1}{\tau_t} = \omega_T \text{ (Transit Frequency)}$$

Where τ_t is the transit time.

Here τ_t is the transit time taken by the minority carriers to cross the base width. The mechanism of transit is both diffusion and drift.

Now let us consider CE BJT at high frequency:

Here β_f (short circuit current gain in CE BJT) is arrived at in exactly the same manner as α_f was arrived at in CB BJT.

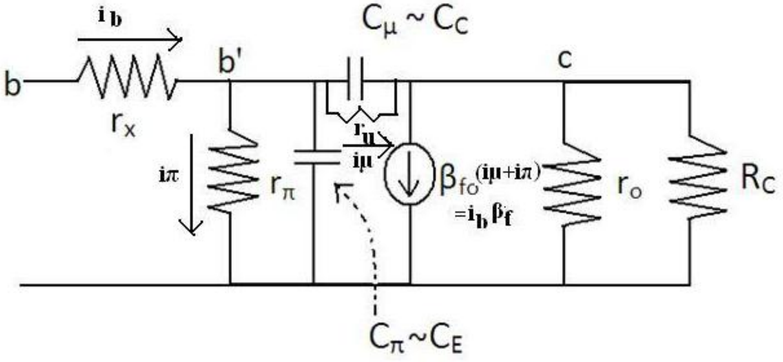


Figure 17. High Frequency Hybrid-pi Model

Figure 17. High Frequency Hybrid-π Model of CE BJT.

$$\beta_f = \textit{Short circuit current gain}$$

$$\beta_f = \frac{\beta_{fo}}{1 + i\,\frac{\omega}{\omega_\beta}}$$

Where ω_β=β cutoff frequency=$\frac{1}{r_\pi(C_\pi+C_\mu)}$

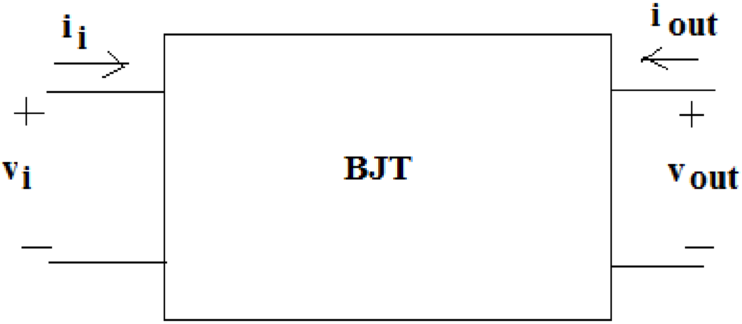
$$= \frac{1}{\frac{\beta_o}{g_m}C_\pi} = \omega_\alpha/(1+\beta_0)$$

Here we neglect C_μ assuming $C_\pi \gg C_\mu$.

SSPD_Chapter 4_ Section 4.1.Physical Significance of hybrid parameters (h-parameters) of BJT.
This Chapter gives physical significance of the four hybrid parameters and on the basis of these four h-parameters the salient differences of CB and CE BJT are brought out.Early Effect or Base Width Modulation and Conductivity Modulation are explained.

4.1. Physical Significance of hybrid parameters (h-parameters) of BJT.

In small signal analysis of solid state circuits, we use h-model for BJT either connected as Common Base or Common Emitter configuration. We treat BJT as a Black Box and we look at open circuit or short circuit parameters at the input and output port as shown in Figure 4.1.1.



$$\begin{bmatrix} v_i \\ i_{out} \end{bmatrix} = \begin{bmatrix} h_i & h_r \\ h_f & h_o \end{bmatrix} \times \begin{bmatrix} i_i \\ v_{out} \end{bmatrix}$$

Figure 4.1.1. Definition of h-parameters of BJT connected as CB or CE configuration.

Definition of h-parameters:

$$\frac{v_i}{i_i}\Big|_{v_{out}=0} = h_i(short\ circuit\ parameter\ input\ impedance) \qquad 4.1.1.$$

Here incremental output voltage is kept at zero to measure input impedance h_i of BJT circuit configuration. Incremental Output Voltage being kept at zero means that output DC Voltage(V_{CB} in case of CB circuit configuration and V_{CE} in case of CE configuration) is kept constant.

$$\frac{v_i}{v_o}\bigg|_{i_{in}=0} = h_r(open\ circuit\ parameter - reverse\ transmission\ factor) \quad 4.1.2.$$

Here incremental input current is kept at zero to measure reverse transmission factor h_r of BJT circuit configuration. Incremental Input Current being kept at zero means that input DC Current(I_E in case of CB circuit configuration and I_B in case of CE configuration) is kept constant.

$$\frac{i_o}{i_i}\bigg|_{v_{out}=0} = h_f(short\ circuit\ current\ gain) \quad 4.1.3.$$

Here incremental output voltage is kept at zero to measure forward short circuit current gain h_f of BJT circuit configuration. Incremental Output Voltage being kept at zero means that output DC Voltage (V_{CB} in case of CB circuit configuration and V_{CE} in case of CE configuration) is kept constant.

$$\frac{i_o}{v_o}\bigg|_{i_{in}=0} = h_o(open\ circuit\ parameter - output\ admittance) \quad 4.1.4.$$

Here incremental input current is kept at zero to measure output admittance h_o of BJT circuit configuration. Incremental Input Current being kept at zero means that input DC Current(I_E in case of CB circuit configuration and I_B in case of CE configuration) is kept constant.

For CB Circuit configuration, subscript b is added and for CE configuration subscript e is added.

Namely for CB configuration the four parameters are:

$h_{ib}, h_{rb}, h_{fb}, h_{ob}$

Namely for CE configuration the four parameters are:

$h_{ie}, h_{re}, h_{fe}, h_{oe}$

The typical parameters for the four h-parameters in CB and CE configuration are tabulated in Table 4.1.1.

Table 4.1.1. Typical values of the four h-parameters in CB and CE configurations.

	Physical significance of h-parameter	**CB**	**CE**
h_i	Input impedance of the given configuration	$h_{ib} = 26\ \Omega$	$h_{ie} = 2.6\text{k}\Omega$
h_r	Reverse transmission	$h_{rb} = 10^{-4}$	$h_{re} = 2.6 \times 10^{-3}$

	factor from output to input		
h_f	Forward short circuit current gain	$h_{if} = -0.99 = \alpha_{fo}$	$h_{if} = 100 = \beta_{fo}$
h_o	Output Admittance of the given configuration	$h_{ob} = 1/(1M\Omega)$	$h_{ie} = 1/(40k\Omega)$

From the output characteristics of CB configuration and CE configuration it is obvious that CB is an ideal current source and CE is a non-ideal current source. A set of vertical I-V characteristics represents a controlled voltage source whereas a set of horizontal I-V characteristics represents a controlled current source. A sloping I-V characteristics shows a current source with finite but large output impedance. In CB family of curves we have a set of horizontal lines with very small slope hence its output impedance at the Quiescent Point is very large. In this case the output impedance is almost 1MΩ or larger. Hence we say that CB Configuration is ideal current source. In case of CE configuration there is a much larger slope which results in a lower output impedance of 40kΩ.

This slope results due to Base-Width Modulation also known as Early Effect. This we will discuss in the next section.

Next we look at the reverse transmission factor. It is one order of magnitude smaller in CB configuration as compared to that in CE configuration. This means that feed-back from output to input is negligible and BJT is an Unilateral Device in case of CB configuration. Whereas there is much larger feedback from output to input in case of CE configuration and CE BJT behave as an non-Unilateral device. Hence RF Amplifiers which are susceptible to *Parasitic Oscillations* prefer CB configuration as compared to CE configuration.

The fact that CB is Unilateral device and CE is non-Unilateral device is evident from Figure 4.1.2. also.

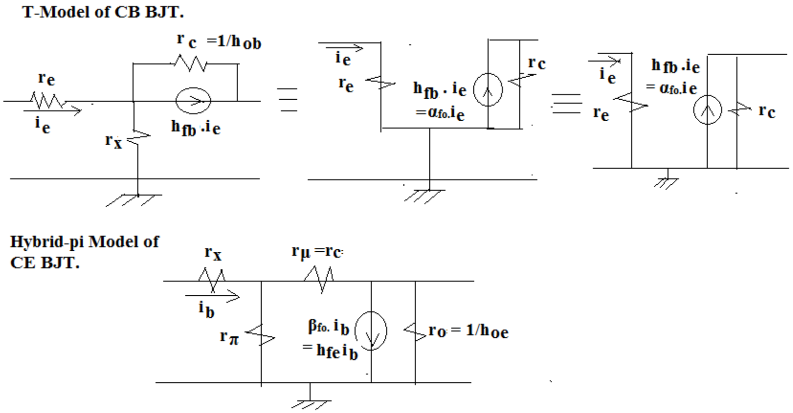


Figure 4.1.2. The incremental T-Model of CB BJT and incremental Pi-Model of CE BJT.

T-Model of CB BJT has two meshes, input mesh and output mesh, which are very weakly interacting through base spreading resistance, r_x , and are completely non-interacting if $r_x = 0$. Hence CB BJT is near Unilateral Device and inspecting the diagram we see that:

$$h_{rb} = \frac{r_x}{r_c} = \frac{100\Omega}{1M\Omega} = 10^{-4} \ \ at\ 1mA = I_C\ current \qquad 4.1.5$$

Hybrid-π Model of CE BJT has two very strongly interacting input and output meshes even if the base spreading resistance is zero. Therefore CE BJT is considered to be non-Unilateral device therefore it is highly susceptible to parasitic oscillations in RF Applications. Hence CE BJT is never utilized for building a RF Amplifier. Inspecting the diagram wew see that:

$$h_{re} = \frac{r_\pi}{r_\mu} = \frac{2.6k}{1M} = 2.6 \times 10^{-3}\ at\ 1mA = \ I_C current \qquad 4.16.$$

Input impedance of CB BJT is only 26Ω whereas input impedance of CE BJT is 2.6kΩ at $I_C = 1mA$.

The short circuit current gain in CB BJT is less than Unity hence called short circuit current transfer ratio:

$$h_{fb} = -0.99 = \alpha_{fo} \ \ at\ low\ frequency \qquad 4.17.$$

The short circuit current gain in CE BJT is $\beta_{fo} \sim 100$. Because of Unity Short Circuit Current Gain in CB BJT, it has a much larger cut-off frequency as compared to that of CE BJT:

$$\omega_{\alpha o} = (1 + \beta_{fo})\omega_{\beta 0} = 2\pi \times \frac{800Mradians}{s} \qquad 4.18.$$

Alpha cut-off frequency or Alpha -3dB frequency = 800MHz and

Beta cut-off frequency = 8 MHz.

4.1.1.1. Early Effect or Base-Width Modulation.

As the output voltage in CB or CE configuration varies it leads to variable reverse bias voltage across BC Junction. This leads to variation in the Neutral Base Region. A narrow base gives a better transistor action and a wide base gives a poor transistor action. As output voltage decreases, neutral base width decreases leading to increase in Collector Current. This gives the slope to the output characteristic.

In CB BJT, transistor action is a weak function of W = width of the neutral base hence in CB BJT there is small slope in I-V curve and output impedance is more than 1MΩ at 1mA collector current.

In CE BJT, transistor action is a strong function of W therefore we obtain a larger slope in I-V curve and output impedance falls to 40kΩ at 1mA Collector Current.

Both these points are evident when we look at the output characteristics of CB and CE BJT.

4.1.1.2. Forward Beta (β_F) fall off at low and high currents.

At moderate currents from 0.1mA to 5mA short circuit current gain is constant but it falls off at very low currents and at very high currents as shown in Figure 4.1.3.. Just as in Diode there is a deviation from the ideal Diode equation at low currents due to recombination in the depletion region in exactly the same way the injection efficiency in BJT falls off at low currents due to recombination in the depletion region. This leads to a deterioration in **βF** at low currents.

Again just as in Diode there is a deviation from ideal Diode equation at high currents due to High Injection Level(HIL) , in the same way there is a conductivity modulation of Base at high currents due to HIL effects. This again leads to the deterioration in injection efficiency and hence in **βF**at high cureents.

In the design section we will show that for a good injection efficiency conductivity of emitter must be much larger than that of the Base. But this condition is not satisfied at high injection level.

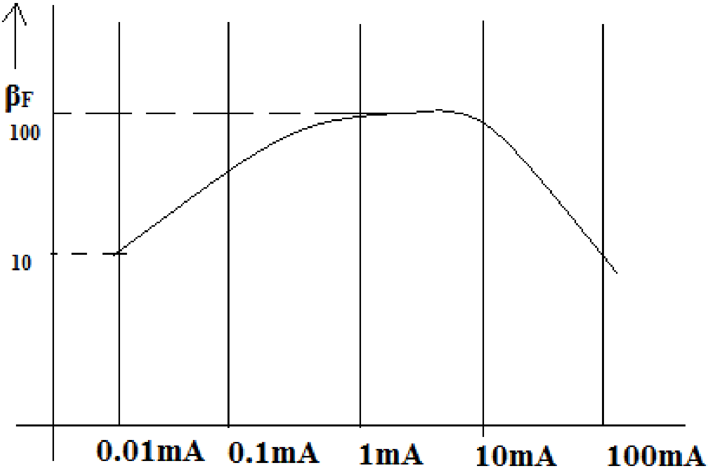


Figure 4.1.3. Fall off of short circuit current gains at low and at high currents.

Chapter 4. Section 4.2. Structure of IC NPN BJT.

Section 4.2 gives the structure, diffusion profile and concentration profile of NPN BJT.

Chapter 4. Section 4.2. Structure of IC NPN BJT.

In Figure 4.2.1. the Cross-sectional view of Integrated Circuit Vertical NPN Transistor is shown.

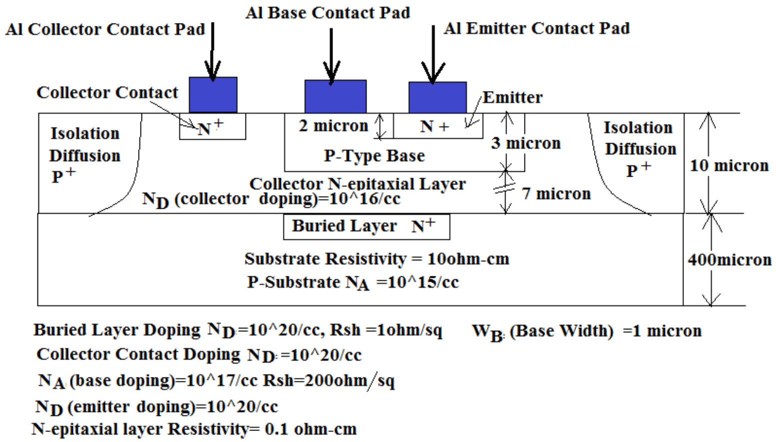

Al Base Contact Pad
Al Collector Contact Pad
Al Emitter Contact Pad
Collector Contact
Emitter
Isolation Diffusion P+
N+
2 micron
N +
3 micron
P-Type Base
Isolation Diffusion P+
10 micron
Collector N-epitaxial Layer
ND (collector doping)=10^16/cc
7 micron
Buried Layer N+
Substrate Resistivity = 10ohm-cm
P-Substrate NA =10^15/cc
400micron
Buried Layer Doping ND=10^20/cc, Rsh =1ohm/sq
WB (Base Width) =1 micron
Collector Contact Doping ND =10^20/cc
NA (base doping)=10^17/cc Rsh=200ohm/sq
ND (emitter doping)=10^20/cc
N-epitaxial layer Resistivity= 0.1 ohm-cm

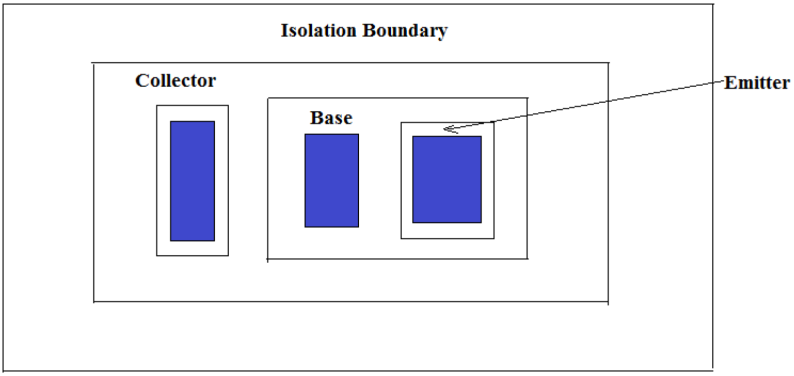
Figure 4.2.1. Cross-sectional View of IC Vertical NPN BJT. with dimensions as in 1970

In 1960, we had micron scale BJT with Base Width(W_B) = 1μm. Today in 2013 we have nano-scale BJT with Base Width = 10nm. In last 50 years with the advancement of IC Technology we have scaled down the Base Width by 1/100 times. This has resulted in improvement of Transit Frequency f_T from 100MHz to 500GHz. This is 5000 times improvement in speed.

Figure 4.2.2. gives the top view of Vertical NPN

Transistor. As seen in Figure 4.2.1. There is 400 micron thick wafer of 300mm diameter. Now we are moving to 450mm diameter wafer for IC fabrication. This is called the substrate and it provides the mechanical strength to the IC fabricated over it. Where ever vertical NPN transistor is situated underneath it 1 micron deep Buried Layer is provided to give a very low resistance path to the collector current. If Buried Layer is not there then collector bulk series resistance will become high leading to large Vsat voltages.

In saturation mode, transistor operates in cut-off mode and saturation mode. In Cut-off Mode transistor gives 5V output which corresponds to ‘1’ bit. In Saturation Mode transistor output is 0.2V and this corresponds to ‘0’ bit. If buried layer is not introduced then there will be considerable drop across series resistance of Collector bulk leading to Vsat = more than 1V. This is not a realistic ‘0’ bit. Hence buried layer is essential in case of Logic Applications.



Blue rectangles are metallic contact pads
Figure 4.2.2. Plan (TOP)View of IC Vertical NPN Transistor with vertical stripe geometry.

In Figure 4.2.3. the concentration profile before and after compensation is shown for IC Vertical NPN Transistor. The substrate is a wafer of 400μm thickness and has uniform doping of 10^{15} per cc which corresponds to a $\rho = 10$ Ω-cm. Over the substrate a 10μm thick N-Type epitaxial layer is grown with a background doping of 10^{16} per cc which corresponds to 0.1 Ω-cm. The actual IC components are fabricated in the epitaxial layer. The substrate is only for providing mechanical strength to the IC. Metallic interconnections (of Al or Cu) are made on the top surface of the epitaxial layer. Hence it is called Planar Technology.

Base diffusion is a long drive-in for several hours from a limited source up to more than 3μm depth. Hence Base Diffusion Profile is a Gaussian Profile.

Emitter Diffusion is a short diffusion from infinite source. Hence Emitter Diffusion is a complementary error function.

After compensation we obtain an Abrupt Junction at EB metallurgical junction and we obtain linearly-graded junction at BC metallurgical junction.

Emitter is heavily N-type doped at 10^{20}/cc. This is equivalent to Sheet Resistance $R_{Sheet} = 1\Omega$/square. In contrast Base is linearly graded from 10^{17}/cc on Emitter side to 0 /cc on Collector side. This corresponds to $R_{Sheet} = 200\Omega$/cc. This large differential in sheet resistance helps achieve Emitter Injection Efficiency of nearly 100%. The linearly graded impurity profile creates an in-built electric field which aids the diffusion of minority carriers across the Base Junction as we will see in the Design Rules.

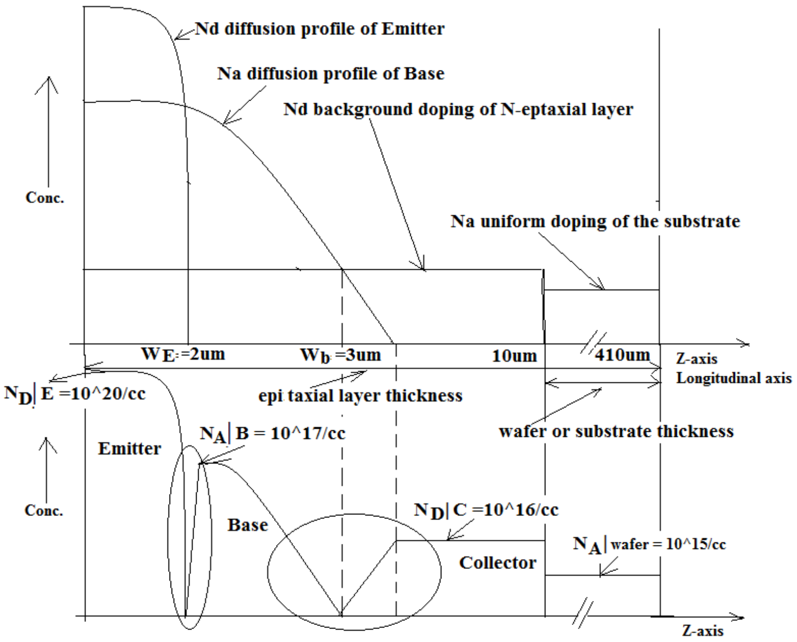
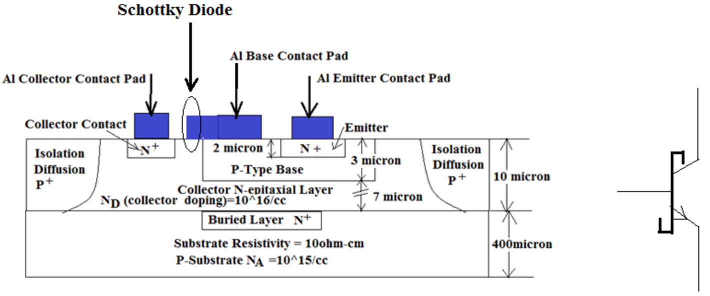

Nd diffusion profile of Emitter
Na diffusion profile of Base
Nd background doping of N-eptaxial layer
Conc.
Na uniform doping of the substrate
WE =2um
Wb =3um
10um
410um
Z-axis
Longitudinal axis
ND| E =10^20/cc
epi taxial layer thickness
wafer or substrate thickness
Emitter
NA| B = 10^17/cc
Conc.
Base
ND| C =10^16/cc
Collector
NA| wafer = 10^15/cc
Z-axis

Figure 4.2.3. (a)Concentration Profile before compensation.
(b) Concentration Profile after compensation.

In Figure 4.2.4. we show the structure of Schottky Clamped NPN Transistor. As shown in the Figure 4.2.4. the Base Metallic Contact is extended upto collector. Metal-N-Type Semiconductor Interface becomes a Schottky Diode.

This modification greatly reduces the switching time of TTL gates. Without Schottky Diodes, when “1” bit is impressed on the Base of a RTL Inverter, the base current drives the Gate Transistor is into ‘DEEP SATURATION’. But in case of Schottky Clamped transistor, even before BC Junction is forward biased, Schottky diode turns ON and shunts the base current as a result Gate Transistor is driven into Shallow Saturation and minority carrier build-up in

Base is limited. Hence Transistor easily switches off. This results into much faster switching of Schottky Clamped TTL gates.



(a) Cross-sectional view of Schottky Clamped NPN BJT

(b)Symbol of Schottky Clamped BJT.

Figure 4.2.4. (a) The structure and (b) the Symbol of Schottky Clamped NPN BJT.

Chapter 4_Section 4.3. Design Rules for fabricating BJT.
Section 4.3 gives the design rules for fabricating BJT of a given Beta and Transit Frequency.

Section 4.3. Design Rules for fabricating BJT.

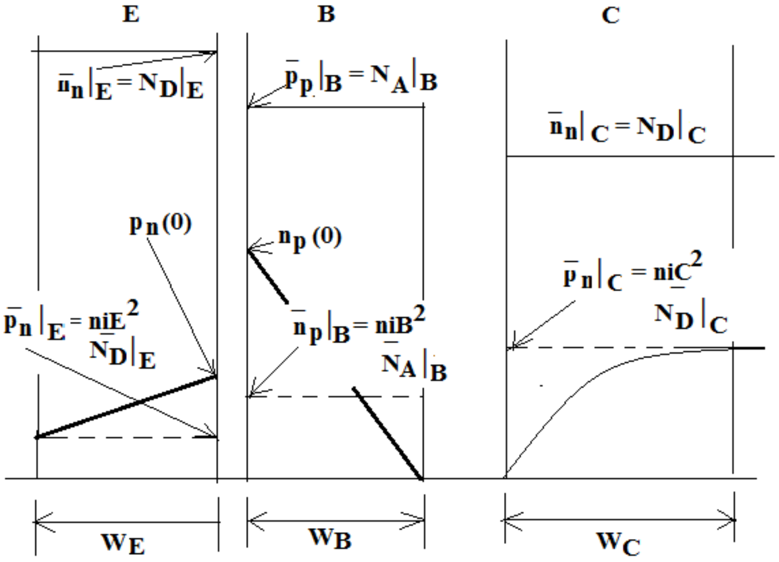


Figure 4.3.1.CarrierConcentrationProfile in Forward Active Mode of IC Vertical NPN Transistor.

In Figure 4.3.1. the carrier concentration profile of Forward Active Mode BJT is revisited. Because of Forward Bias at EB Junction, electrons are injected from E to B and holes are injected from B to E.

In micro scale devices: base current had two parts = $I_{B1} + I_{BB}$.

$$I_{BB} = \frac{Q_B}{\tau_{nB}} \qquad and\ I_{B1} = current\ injected\ into\ the\ Base =$$

$$Diffusion\ Current\ in\ the\ Emitter = qA_E D_{pE} \frac{p_n(0) - p_n^-}{W_E} = \frac{qA_E D_{pE}}{W_E} p_n^- \left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 1\right] \qquad 4.3.1.$$

Therefore :

$$I_{B1} =$$

$$\frac{qA_E D_{pE}}{W_E} p_n^- \left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 1\right] = \frac{qA_E D_{pE} n_{iE}^2}{W_E N_{DE}} \left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 1\right] \qquad 4.3.2.$$

In nano- scale devices, I_{BB} (the recombination current in Base) is negligible because of 10nm narrow Base. Hence the base current I_B = Diffusion Current in Emitter = I_{Ep}.

Therefore:

$$I_B = I_{B1} = \frac{qA_E D_{pE} n_{iE}^2}{W_E N_{DE}} \left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 1\right] = qA_E \frac{D_{pE}}{W_E} \frac{n_{iE}^2}{N_{DE}} \left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 1\right] \qquad 4.3.3.$$

Let us define a quantity Gummel Number:

$$G_E = \int_0^{W_E} \frac{n_i^2}{n_{iE}^2} \frac{n_n(z)}{D_{pE}} dz \qquad 4.3.4.$$

This is known as Emitter Gummel Number and it has the unusual dimension of sec/cm^4.

At low level injection $n_n(z) = N_{DE}$ in Emitter. Let $n_i^2 = n_{iE}^2$ and let D_{pE} be constant throughout the Emitter Bulk then:

$$G_E = \frac{1}{D_{pE}} \int_0^{W_E} N_{DE} dz = \frac{N_{DE} W_E}{D_{pE}}$$
$$= \frac{\textit{total dopent atoms in Emitter Bulk per unit area}}{D_{pE}} \qquad 4.3.5.$$

Eq.4.3.5. simplifies the Base Current Model and it contains all the subtleties of transistor design that affect I_B :

1. if we change the emitter material and its band gap, it is accounted for by n_{iE}^2;
2. If we have non-constant D_{pE} then it will be included in the Integral;
3. If the majority carrier is not constant it will be accounted by taking non-uniform emitter dopent namely $N_{DE}(z)$;
4. Even high level injection will be take care of by taking the high injection effect into Emitter Gummel Number.

By substituting Eq.4.3.5 in Eq.4.3.3. we get:

$$I_B = qA_E \frac{n_i^2}{G_E}\left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 1\right] \qquad 4.3.6.$$

In exactly the same manner, collector current I_C can be determined.

$$I_C =$$

$$\frac{qA_E D_{nB}}{W_B} n_p^- \left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 0\right] = \frac{qA_E D_{nB} n_{iB}^2}{W_B N_{AB}}\left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 0\right] \qquad 4.3.7.$$

Defining Base Gummel Number as follows:

$$G_B = \int_0^{W_B} \frac{n_i^2}{n_{iB}^2}\frac{p_p(z)}{D_{nB}}dz \qquad 4.3.8.$$

Taking the special case of low level injection $p_p(z) = N_{AB}$ in Emitter , $n_i^2 = n_{iB}^2$ and D_{nB} be constant throughout the Emitter Bulk we get:

$$G_B = \frac{1}{D_{nB}}\int_0^{W_B} N_{AB}\,dz = \frac{N_{AB}W_B}{D_{nB}} = \frac{\mathit{total\ dopent\ atoms\ in\ neutral\ Base\ Bulk\ per\ unit\ area}}{D_{nB}} \qquad 4.3.9.$$

Therefore

$$I_C = qA_E\frac{n_i^2}{G_B}\left[Exp\left(\frac{V_{BE}}{V_{Th}}\right) - 0\right] \qquad 4.3.10.$$

Section 4.3.1. DC Current Gain of CE BJT.

Short Circuit CE BJT Current Gain = I_C / I_B = β_F ________________________4.3.11.

Using Eq.4.3.10 and Eq.4.3.6. we get :

Short Circuit CE BJT Current Gain = β_F = G_E/G_B

Therefore

$$\beta_F = \frac{G_E}{G_B} = \frac{qA_E D_{nB} n_{iB}^2}{W_B N_{AB}} \times \frac{1}{qA_E\frac{D_{pE}}{W_E}\frac{n_{iE}^2}{N_{DE}}} = \frac{W_E N_{DE} D_{nB} n_{iB}^2}{W_B N_{AB} D_{pE} n_{iE}^2} \qquad 4.3.12$$

From Eq.4.3.12.in order to obtain large β_F = 100, we must have high Emitter Doping and light Base

Doping but light Base Doping will lead to large Base Spreading Resistance r_x which leads to deleterious effect on high current and at high frequency operation. So we strike a balance by keeping Emitter doping at 10^{20}/cc and Base Doping at 10^{18}/cc .

To obtain high Emitter Injection Efficiency γ we must examine the expression γ.

Injection Efficiency had been defined as:

$$\gamma = \frac{I_{En}}{I_E} = \frac{I_C}{I_C + I_B} = \frac{1}{1+\frac{I_B}{I_C}} = \frac{1}{1+\frac{G_B}{G_E}}; \qquad 4.3.13$$

If $\beta_F = 100$ then $\alpha_F = 0.99$.

Since $\alpha_F = I_C/I_E = \beta_F / (1+ \beta_F)$ and $\beta_F = \alpha_F /(1-\alpha_F)$

Injection Efficiency $= \gamma = 1/(1+1/\beta_F) = \alpha_F$;

Section 4.3.2. Emitter Band Gap Narrowing.

Because of heavy doping effect there is a Band Gap Narrowing by ΔV. This results in a reduction of Short Circuit CE BJT Current Gain.

We know that :

$$n_i^2 = N_C N_V Exp\left(-\frac{E_g}{kT}\right) \qquad 4.3.14$$

If there is Band Gap narrowing by ΔV then

Eq.4.3.14 gets modified to:

$$n_{iE}^2 = N_C N_V Exp\left(-\frac{E_g - \Delta E}{kT}\right) \qquad 4.3.15$$

Dividing Eq.4.3.15 by Eq.4.3.14 we get :

$$n_{iE}^2 = n_i^2 Exp\left(\frac{\Delta E}{kT}\right) \qquad 4.3.16$$

Substituting Eq.4.3.16 in Eq.4.3.12 we get:

$$\beta_F = \frac{G_E}{G_B} = \frac{W_E N_{DE} D_{nB} n_{iB}^2}{W_B N_{AB} D_{pE} n_{iE}^2} = \frac{W_E N_{DE} D_{nB}}{W_B N_{AB} D_{pE} Exp\left(\frac{\Delta E}{kT}\right)} \qquad 4.3.17.$$

Typically ΔE = 0, 50meV, 95meV at less than 1018/cc , 1019/cc , 1020/cc , respectively.

Thermal energy is 25meV. Therefore at 1019/cc , 1020/cc , βF will fall by 1/7.39 and 1/36.6 respectively.

Section 4.3.3. Base Transit time and Charge Storage in Base.

Referring to Figure 4.3.1. the charge stored in the Base:

$$Q_{FB} = \frac{q A_E W_B n_{p0}}{2} \qquad 4.3.18$$

And Collector Current is:

$$I_C = qA_E D_{nB} \frac{n_{p0}}{W_B} \quad 4.3.19.$$

We define Q_{FB}/I_C = Forward Transit Time = τ_{FB} .

Dividing Eq.4.3.18 by Eq.4.3.19 we get:

$$\tau_{FB} = \frac{W_B^2}{2D_{nB}} \quad 4.3.20$$

We purposely make Base doping density linearly graded. This gives rise a built-in electric field which opposes Majority Carriers but support the diffusion of Minority Carriers from higher to lighter density. This electric field shortens the Transit Time in forward direction. Hence :

$$\tau_{FB} = \frac{W_B^2}{KD_{nB}} \quad where\ K > 2 \quad 4.3.21$$

Therefore we get a better performance than expected. Therefore linearly graded base Transistors are referred to as Drift Transistor.

Typically Effective Base Width is 70nm and D_{nB} = 10 cm^2/s. This gives a transit time by Eq.4.3.20 as 2.5ps.

Therefore Transit angular frequency = ω_T = 1/2.5ps = 400Grad/s

Transit Frequency = 400/(2π) = 64GHz.

To date we have achieved 500GHz with Base Width of 10nm.

Chapter 4. Section 4.3.supplement-Solved Problems on Design of BJT.
Section 4.3 supplement gives some solved problems in designing BJT.

Chapter 4. Section 4.3.supplement-Solved Problems on Design of BJT.

Problem 1. A BJT has I_C = 1mA and I_B = 10µA. What are α_F , β_F and I_E ?

Solution:

$$I_E = I_C + I_B = 1mA + 10\mu A = 1.01mA.$$

$$\beta_F = \frac{I_C}{I_B} = \frac{1mA}{10\mu A} = 100 \ and \ \alpha_F = \frac{I_C}{I_E} = \frac{1mA}{1.01mA} = 0.9901$$

You can check for consistency that :

$$\alpha_F = \frac{\beta_F}{1+\beta_F} = 0.9901 \ \ and \ \beta_F = \frac{\alpha_F}{1-\alpha_F} = 100.$$

Problem 2. Heavy Emitter Doping effect on Short Circuit CE BJT Current Gain. In a BJT D_{nB} (diffusion coefficient of electron in Base) = 3 D_{pE} (diffusion coefficient of holes in Emitter); Emitter Bulk Width = W_E = 3W_B; Base Doping is uniform = N_B = 10^{18}/cc; $n_{iB}^2 = n_i^2$;

Determine β_F for (a) $N_E = 10^{19}$/cc; and (b) $N_E = 10^{20}$/cc ? [include band gap narrowing due to heavy emitter doping]

Solution: Rewriting Equation 4.3.17 we get:

$$\beta_F = \frac{D_{nB}W_E}{D_{pE}W_B} \times \frac{N_E}{N_B} \times \frac{1}{Exp[\frac{\Delta E_g}{kT}]} \tag{1}$$

Case (a) At $N_E = 10^{19}$/cc , Band-gap narrowing = 50meV

Therefore:

$$\beta_F = 9 \times \frac{10^{19}}{10^{18}} \times \frac{1}{6.84} = 13.15.$$

Case (b) At $N_E = 10^{20}$/cc ,, Band-gap narrowing = 95meV

Therefore β_F = 900/38.6 = 23.3.

Band-gap narrowing has a very adverse effect. If it was not considered then S.C.Current Gain would have been 90 and 900 but now it has been reduced to 13 and 23.

Problem 3. Determine the forward Base Transit Time if W_B = 70nm and D_{nB} = 10cm^2/s ? Also determine the improvement in forward Transit Time if W_B = 10nm and D_{nB} = 10cm^2/s ?

$$\tau_{FB} = \frac{W_B^2}{2D_{nB}} = 2.45\ ps\ and\ 50\ femtos\ for\ 10nm\ Bae\ Width.$$

This corresponds to Transit angular Frequency or Unity Current Gain angular BW:

$$\omega_T = \frac{1}{\tau_{FB}} = \frac{408Grads}{s}\ for\ 70nm\ Base\ Width\ and\ 20\ Tera - rads/s$$

This corresponds to :

$$f_T = \frac{\omega_T}{2\pi} = 65\ GHz\ for\ 70nm\ and\ 3.18THz\ for\ 10nm\ Base\ Width.$$

AE Lecture 8_Part A_MOSFET & JFET continued
AE Lecture 8_Part A gave the quantitative treatment of MOSFET. The continuation of Part A gives the quantitative treatment of JFET.

AE_Lecture 8_MOSFET & JFET_Part A continued.

In Part A of Lecture 8 we have seen the quantitative formulation of Enhancement and Depletion type MOSFET in Triode as well as in Pentode region. In this lecture, the continuation of Lecture 8_Part A, we will study the quantitative formulation of nJFET in ohmic and saturation region also known as triode and pentode region respectively.

Enhancement Type MOSFET was termed as Normally-Off device since with Gate-Voltage zero we have no current.

Similarly Depletion Type MOSFET is termed as Normally-On device since here even under Gate Voltage zero we do have a drain current because in NMOS(D), n-channel has been ion-implanted at fabrication stage hence it conducts at zero Gate Voltage.

In exactly the same fashion, JFET has a conducting channel at Gate Voltage zero. Hence it conducts under zero Gate Voltage therefore it is called Normally – On device. It is a depletion-mode device.

The physical structure of nJFET in a simplified manner is given in Figure 1.

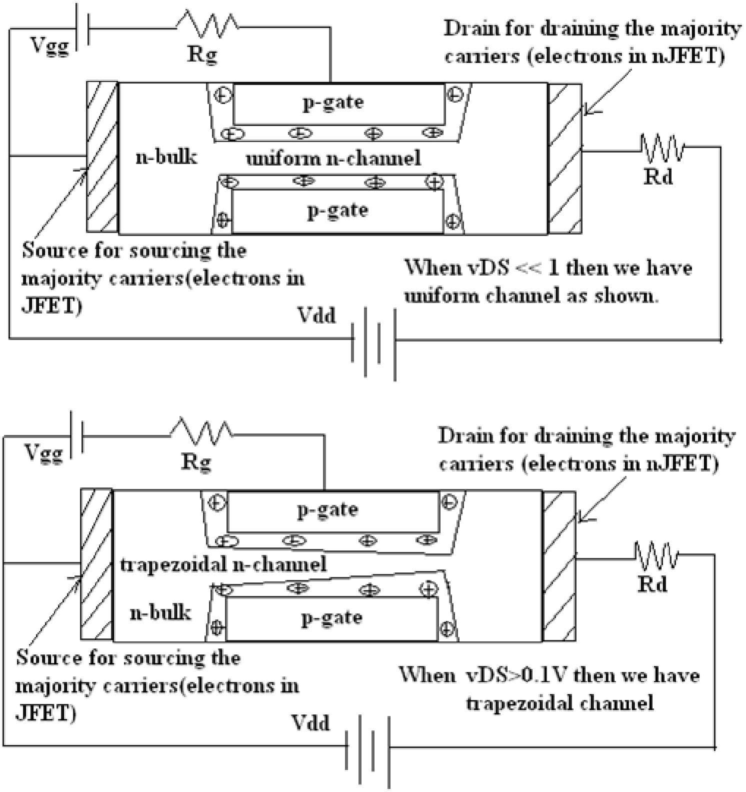

Vgg
Rg
Drain for draining the majority carriers (electrons in nJFET)
p-gate
n-bulk
uniform n-channel
p-gate
Rd
Source for sourcing the majority carriers(electrons in JFET)
When vDS << 1 then we have uniform channel as shown.
Vdd
Vgg
Rg
Drain for draining the majority carriers (electrons in nJFET)
p-gate
trapezoidal n-channel
n-bulk
p-gate
Rd
Source for sourcing the majority carriers(electrons in JFET)
When vDS>0.1V then we have trapezoidal channel
Vdd

Figure 1. Physical structure of nJFET is given for negligible Drain-to-Source biasing and significant Drain-to-Source biasing respectively.

For negligible Drain-to-Source biasing, we have uniform n-channel as shown in the upper diagram. In this biasing condition, nJFET is operating in ohmic region or in Triode region.

For $V_{DS} > 0.1$ V, we have trapezoidal n-channelas shown in the lower diagram. The device is still in Triode region but its ohmic behavior is non-linear.

When $V_{DS} = (V_{GS} – V_P)$ then the n-channel gets pinched off near the Drain and current saturates at that point. Now the device has entered saturation region. For $V_{DS} > (V_{GS} – V_P)$, the current remains constant and device is said to be in Pentode region.

In Ohmic Region:

$i_D = K_n [\ 2(v_{GS} – V_P)\ v_{DS} – v_{DS}^2]$ for $0 < v_{DS} \leq ((v_{GS} – V_P)$……………………….1

For small values of v_{DS} ,

$i_D = K_n [\ 2(v_{GS} – V_P)\ v_{DS}]$ ……………………………………………………………2

At $v_{DS}^* = (V_{GS} – V_P)$, device enters the saturation region since channel is pinched off near the drain.

In saturation region, irrespective of v_{DS} current becomes constant at the value decided by $v_{DS}^* = (V_{GS} – V_P)$. Substituting $v_{DS}^* = (V_{GS} – V_P)$ in Equation 1, we get:

$I_D = K_n (V_{GS} – V_P)^2$ for $v_{DS} \geq v_{DS}^*$ ……………………………………….3

At $V_{GS} = 0$, we have $I_D = K_n (-V_P)^2 =$ this denoted by I_{DSS} ;

Hence

$$K_n = \frac{I_{DSS}}{V_P^2}$$

……………………………………………………………………………

Typical values of Pinch-off voltage $V_P = 0$ to -25V and typical values of I_{DSS} is from 10μA to 10A.

For a given value of I_D , Equation 3 gives two values of V_{GS} . Only one value is physically tenable. It must lie in the range:

…………………………………$V_P < V_{GS} < 0$

The line of demarcation or the pinch-off locus is a parabola. By substituting

$v_{DS^*} = (V_{GS} – V_P)$ in Equation 3 we get:

$I_D = K_n (V_{GS} – V_P)^2 = I_D = K_n (V_{DS^*})^2$ which is a parabola and is defined as pinch off parabola.

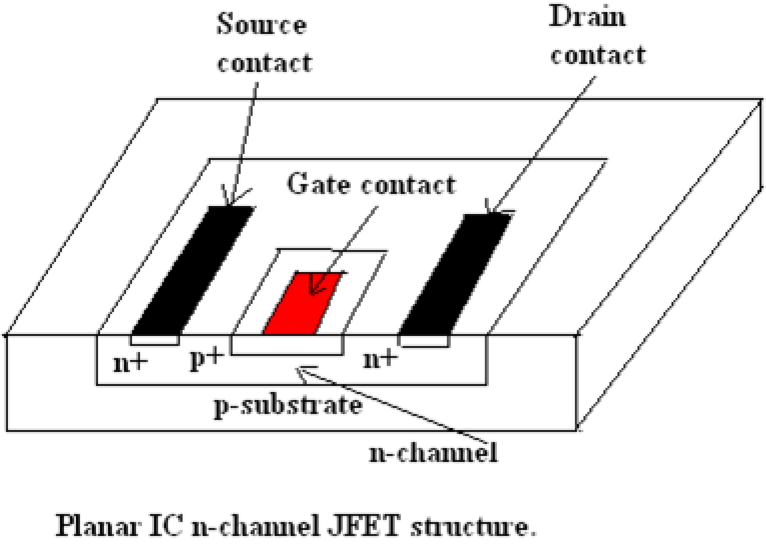


Figure 2. IC version of JFET.

AE_Lecture8_MOSFET & JFET_PartA
AE_Lecture8_PartA describes the structure and static characteristics of MOSFET.

AE_Lecture8_PartA Static characteristics of FIELD EFFECT TRANSISTORS

A Field Effect Transistor is like a pentode. It is an analogue of pentode. Both are Voltage Controlled Current Sources.

FET is of two types JFET and MOSFET.

JFET is n channel FET known as nJFET and p channel FET known as pJFET.

Similarly MOS is n channel NMOS and p channel PMOS.

MOS can be enhancement mode which is normally-off or depletion mode normally-on device.

So NMOS can be (E)NMOS and (D)NMOS.

PMOS can be (E)PMOS and (D)PMOS.

The following is the circuit diagram of an Amplifier using N-channel JFET:

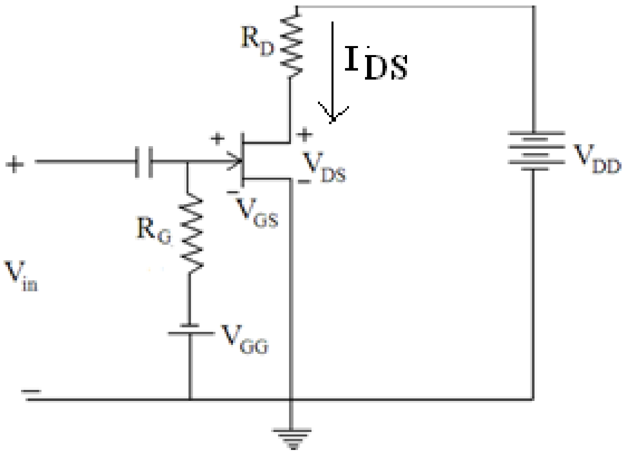


Figure 1. Circuit Set-up for measuring the Static Charactristics of nJFET.

Measurement of characteristics of N-channel JFET

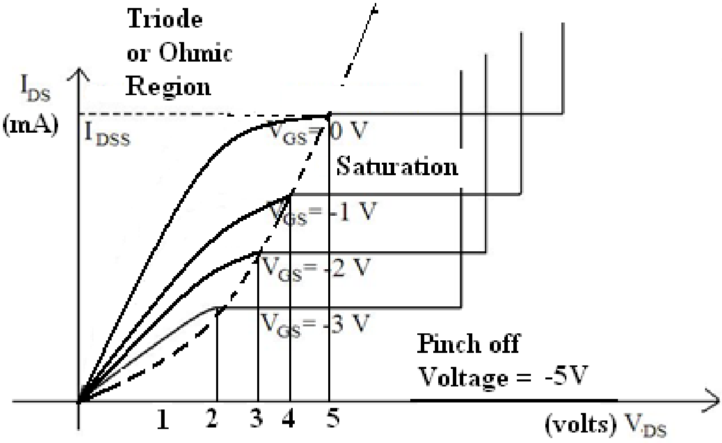


Figure 2. Static Output Characteristics of nJFET.

Figure 2 gives the dc output characteristics nJFET.

Note that Gate Voltage is always kept negative so that the gate current is always zero. We indicate the Gate-Drain breakdown also. As gate voltage becomes more negative , Avalanche Breakdown of Gate to Drain Junction takes place earlier.When $(V_{DS} – V_{GS})$ exceeds V_P (Vpinchoff), the channel gets pinched off towards the Drain End and current becomes constant. This is saturation region. The dotted parabola seperates the Ohmic region or Triode Region from Saturation Region or Pentode Region. Here the pinch off voltage is -5V.

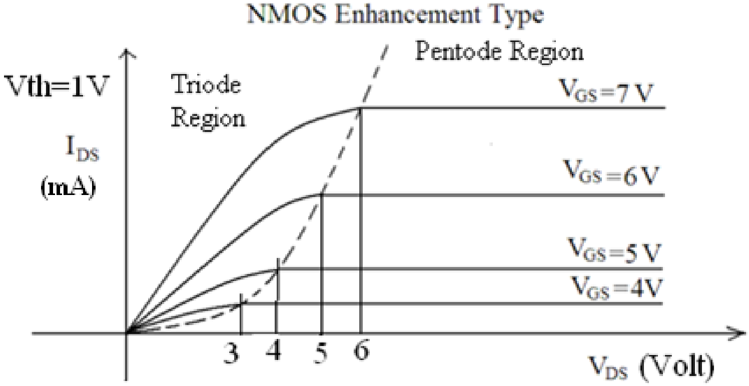


Figure 3. Static Output Characteristics of (E)NMOS.

Figure 3 gives the output characteristics of (E)NMOS with a Threshold Voltage of 1V. Therefore when V_{DS} increases and becomes equal to $(V_{GS} - V_{Th})$ the channel pinches off near the drain end and current saturates.. As V_{DS} assumes larger values the pinched off region becomes wider, $(V_{GS} - V_{Th})$ drops across the channel and excess part of V_{DS} drops across the pinched off region. Since there is a constant voltage dropping across the channel and since the channel is of constant length hence I_{DS} is constant. But in real life there is a shortening of channel length leading to a slope of the saturated current. This channel length modulation is responsible for a finite output impedance r_{ds} of the voltage controlled current source modeling (E)NMOS under incremental condition.

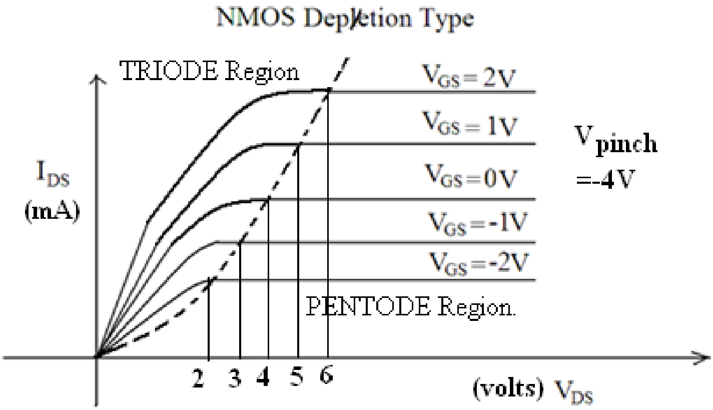


Figure 4.Static Output Characteristics of (D)NMOS.

In Figure 4 we give the output characteristics of (D)NMOS. Depletion type device is normally-on device hence it gives a characteristic for positive gate voltage, zero gate voltage and negative voltage. In (D)NMOS when gate voltage becomes equal to Pinch-off voltage then (D)NMOS turns off. Here pinch off voltage is -4V.

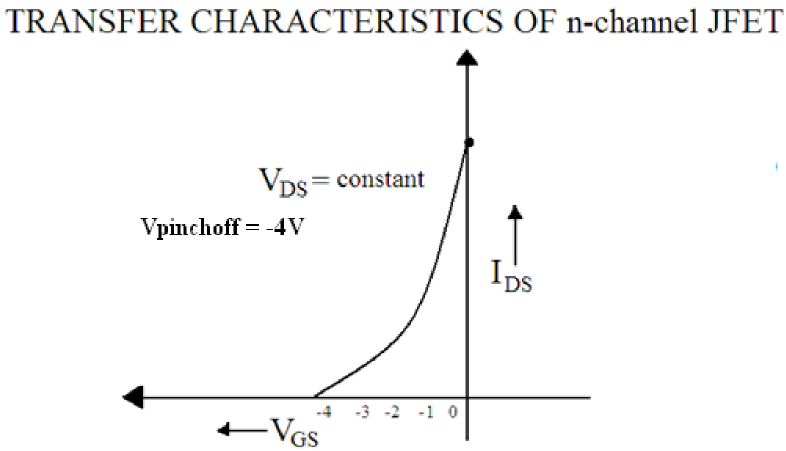


Figure 5. Transfer characteristics of nJFET.

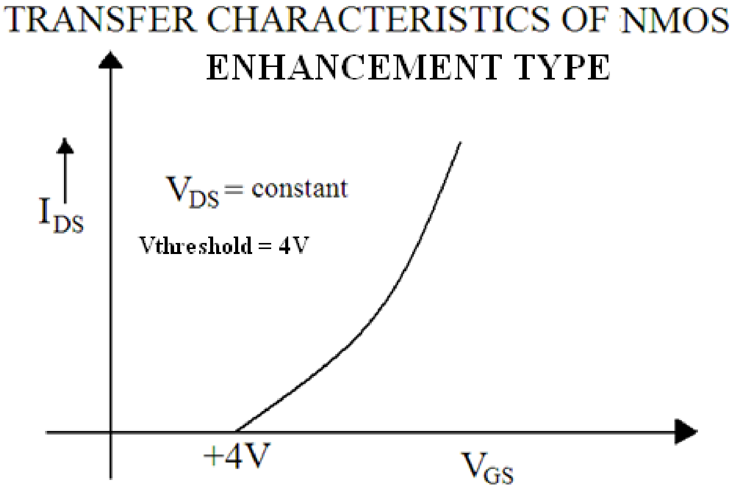


Figure 6. Transfer characteristics of (E)NMOS which is normally off device.

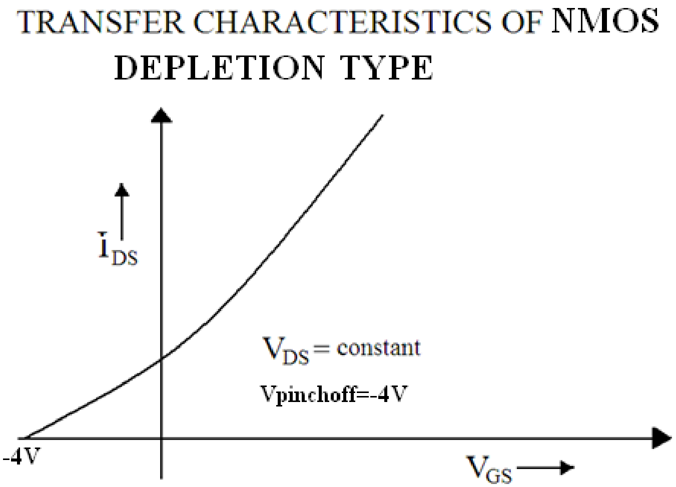

TRANSFER CHARACTERISTICS OF NMOS DEPLETION TYPE

Figure 7. Transfer Characteristics of (D)NMOS which is normally on device.

Comparing the transfer characteristics of nJFET, (E)NMOS and (D)NMOS as given in Figure 5, 6 and 7 we can say that (D)NMOS is most flexible device from design point of view as it accepts both positive and negative voltages as the gate voltage.

NMOS Transistor (Pentode and Triode Region)

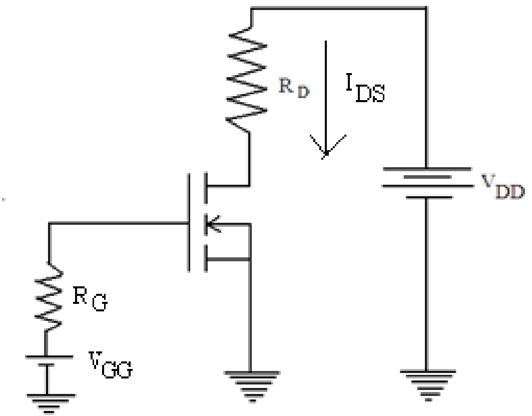


Figure 8. Circuit Set-up to study the Triode and Pentode Region of (E)NMOS.

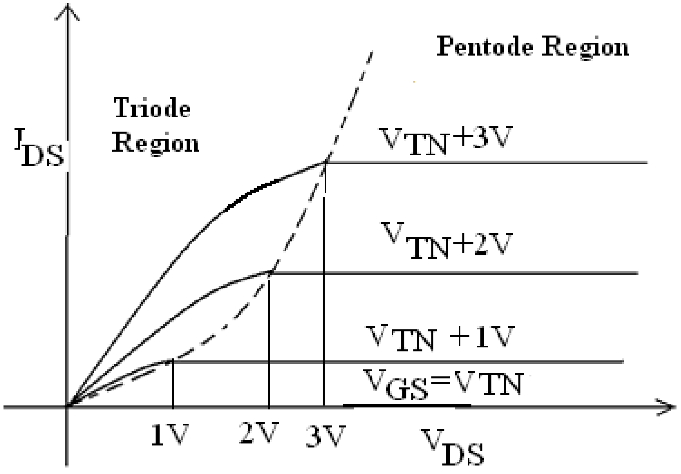


Figure 9. Output Characteristics of (E)NMOS demarcating the TRIODE and PENTODE region.

TRIODE REGION:

$$i_D = K_n\left(v_{GS} - V_{TN} - \frac{(v_{DS})}{2}\right)(v_{DS})$$

____(1)

Kn = Kn'multiplied_by

$$\left(\frac{W}{L}\right)$$

Kn' = μnCox″ ………..Kn’ ia a transconductance parameter which is fixed for a given technology and cannot be changed by the circuit designer. Whereas Kn = Kn’ × (W/L) = this is also a transconductance parameter but it includes the aspect ration (W/L) of the given MOSFET. The aspect ratio gives the geometry of the device and this is under control of a circuit designer . This second transconductance parameter can be controlled when he is generating the masks for a given circuit or system during ASIC(Application Specefic IC) or SOC(System –on-Chip) design. W is the width of the gate and L is the length of the gate or the length of the channel.

$$\text{Cox}'' = \frac{\epsilon_{ox}}{T_{ox}}\left(\frac{F}{cm^2}\right)$$

$$\epsilon_{ox} = oxide\ permittivity\left(\frac{F}{cm}\right) = 3.9 \times 8.854 \times 10^{-14}\,\frac{F}{cm}$$

$$T_{ox} = oxide\ thickness$$

Triode Region exists as long as

$$(V_{GS} - V_{DS}) \geq V_{TN}$$

.

Channel should not pinch off on the drain side.

PENTODE REGION

At

$$V_{DS}^{*} = V_{GS} - V_{TN}$$

________(2)

channel pinches off at drain end. From here onwards current becomes constant or current saturates.

Substituting (2) in (1),

$$I_D = K_n \left(V_{GS} - V_{TN} - \frac{(V_{GS} - V_{TN})}{2} \right) (V_{GS} - V_{TN})$$

$$I_D = K_n \frac{(V_{GS} - V_{TN})^2}{2}$$

$$I_D = K_n^{\square} \frac{W}{L} \frac{(V_{GS} - V_{TN})^2}{2}$$

$$\therefore I_D = K_n^{\square} \frac{W}{L} (V_{GS} - V_{TN})^2$$

$$I_D = \mu_n C''_{ox} \frac{W}{L} (V_{GS} - V_{TN})^2$$

Quadratic Law

$$I_D = \left[C''_{ox} W \frac{(V_{GS} - V_{TN})}{2} \right] \left[\mu_n \frac{(V_{GS} - V_{TN})}{L} \right]$$

Where:

$C''_{ox} W \frac{(V_{GS} - V_{TN})}{2}$ *is the average electron charge per unit length in the inversion layer*

And $\mu_n \frac{(V_{GS} - V_{TN})}{L}$ *is the drift velocity in the channel*

Pinch Off locus is parabolic

$$\because I_D = \frac{K_n (V^*_{DS})^2}{2}$$

For any V_{GS}, at $V^*_{DS} = V_{GS} - V_{TN}$ the saturation sets in as the channel gets pinched off at that point. Beyond that point: $V_{DS} = V^*_{DS} + \Delta V$ and V^*_{DS} drops across the channel and ΔV drops across the depleted region towards the drain end as shown in Figure 11. Since the voltage across the channel is constant at V^*_{DS} and since the channel is approximately constant (it shortens slightly as Drain to Source Voltage increases) for practical purposes we assume that the drain current is constant and hence it has saturated. But in practice due to channel length shortening, there is slight increase in drain current with drain to source voltage. Hence output characteristics does have a slope. At higher currents the slope is higher.

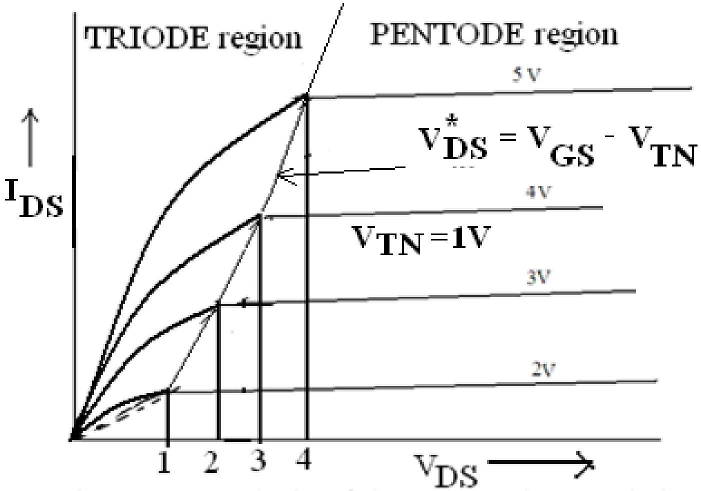


Figure 10. Analysis of the output characteristics of (E)NMOS for demarcating the TRIODE and PENTODE regions.

$V_{TN} = 1V$

$K_n = 25\ \mu A/V^2$

$$i_D = K_n \frac{(V_{GS} - V_{TN})^2}{2}$$

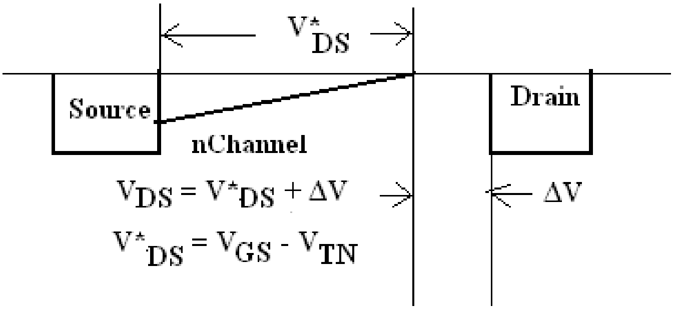


Figure 11. Physical mechanism of current saturation after the channel gets pinched off at drain end. Drain Current continues to flow evn though the channel is interrupted. At the drain end the electron manages to hop across the depletion region by Ballistic Mechanism.

N channel JET has identical formulation except threshold voltage is replaced by pinch off voltage.

AE_Lecture8_PartB_Incremental Model of FET

AE_Lecture8_PartB describes the dynamic characteristics of FET. It models it as Voltage Controlled Current Source. The Figure of Merit is defined.

AE_Lecture8_Incremental Model of FET.

FET behaves as Voltage Controlled Current Source. Hence its model is as given in Figure 1.

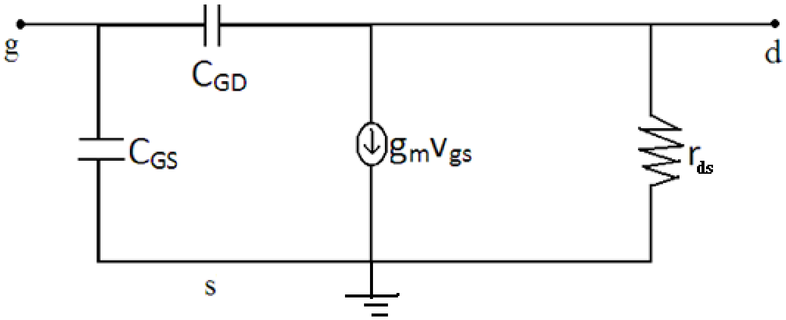


Figure 1. The incremental model of MOSFET.

At low frequencies Capacitances can be considered to be open circuit. And the Incremental Model at low Frequencies will have infinite impedance as shown in Figure 2.

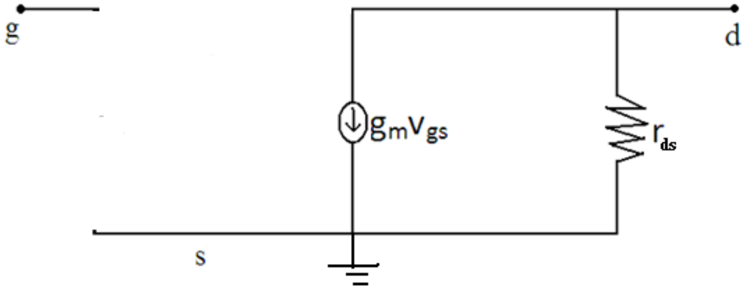


Figure 2. Low Frequency Incremental Model of MOSFET.

At low frequencies, input current is zero and short circuit current gain is INFINITY. But at high frequencies gate to source Capacitance due to overlap Capacitance (overlap between source and gate) and Gate Oxide Capacitance and gate to drain Capacitance again due to overlap Capacitance (overlap between drain and gate) and Gate Oxide Capacitance draw considerable currents. Hence Short Circuit Current Gain is finite and falls at the rate of 20dB/decade.

In the low frequency model, we have transconductance (g_m) and output impedance of the current source (r_{ds}).

Derivation of Transconductance.

$$\frac{di_D}{dV_{GS}}$$

=

$$\frac{(K_n)(2)}{2}(V_{GS} - V_{TN})$$

___(1)

$$g_m = (K_n)(V_{GS} - V_{TN})$$

$$But\ V_{GS} - V_{TN} = \left(\frac{2I_D}{K_n}\right) \times \frac{1}{(V_{GS} - V_{TN})}$$

$$\therefore\ g_m = \frac{(K_n)(2I_D)}{K_n(V_{GS} - V_{TN})}$$

$$g_m = \frac{(2I_D)}{(V_{GS} - V_{TN})}$$

1.

Due to channel length modulation the family of curves are not horizontal but have a finite slope.as shown in Figure 3. To account for this ,channel-length modulation parameter λ is introduced.

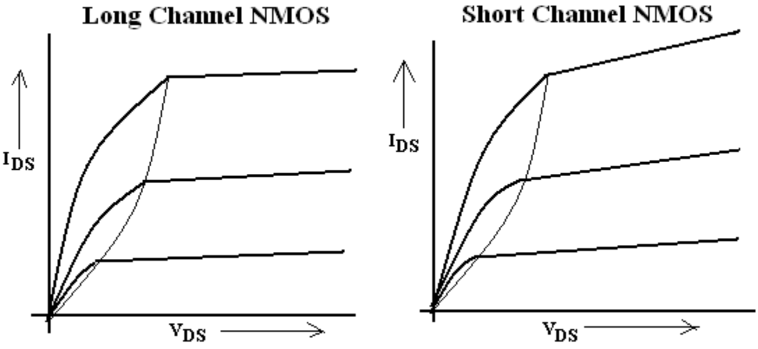


Figure 3. Effect of Channel Length Modulation on the SLOPE.

In long channel NMOS(with channel length of 10μm) of bygone MSI(Medium Scale Integration) and LSI(Large Scale Integration) era, the slope was minimal. But as packing density increased to 40 million transistors, lateral scaling reduced the channel length to 100nm. In short channel NMOS channel length modulation is very pronounced as seen in Figure 3.

$$\therefore I_D = \frac{K_n}{2}(V_{GS} - V_{TN})^2(1 + \lambda V_{DS})$$

$$0.001V^{-1} \leq \lambda \leq 0.1V^{-1}$$

$$\left[\frac{\partial i_D}{\partial v_{ds}}\right] at\ constant\ V_{GS} = \frac{K_n}{2}(V_{GS} - V_{TN})^2(0 + \lambda)$$

$=$

$$\lambda \frac{K_n}{2}(V_{GS} - V_{TN})^2$$

We know that

$$\frac{K_n}{2}(V_{GS} - V_{TN})^2 = \frac{I_D}{1+\lambda V_{DS}}$$

$$\therefore \frac{1}{r_{ds}} = \frac{\lambda I_D}{1+\lambda V_{DS}}$$

=

$$\frac{I_D}{\left(\frac{1}{\lambda}+V_{DS}\right)}$$

2.

Amlification Factor = μ = $g_m \times r_0$ = the maximum gain which can be achieved from a Common Source or Common Emitter Amplifier.

Combining Eq 1 and Eq 2 we get the following:

Amplification Factor of CS MOS Amplifier = [

$$\frac{\frac{1}{\lambda}+V_{DS}}{I_{DS}}$$

]

Amplification Factor of CE BJT Amplifier = [

$$\frac{V_A+V_{CE}}{V_T}]$$

V_A is the Early Voltage which was discussed in Lecture 3 in connection with the Static Output characteristics of CE BJT.

As seen above the amplification factor of CE BJT is independent of Quiescent Collector Current. From the Table we see that it is only 3400. But my research on CE BJT is showing that a much larger gain can be realized.

In the following Table we make a comparative study of BJT and FET.

FET	BJT
Geometry dependent parameters	Geometry independent parameters
$\left(\frac{W}{L}\right)$	
μ(amplification factor	μ(amplification factor is

varies inversely with I_D)	independent of I_D)
$V_{DS} = 10V, K_n = 1 \frac{mA}{V^2}, \lambda = 0.0133\ V^{-1}$	$\beta_o = 100, V_A = 75V, V_{CE} = 10V$
gm = $\left(\frac{I_Q}{V_T}\right)$ $\sqrt{2K_n I_D}$	gm =
Small signal approximation holds good for values: $v_{be} \ll V_T (26mV)$ $v_{gs} \ll 2(V_{GS} - \lvert V_{TN} \rvert)^2$ $v_{be} \approx 0.2V_T$ $v_{gs} \approx 0.2(V_{GS} - V_{TN})^2$ $v_{gs} < 15mV$	Small signal approximation holds good for values:
Much higher packing density because no isolation diffusion required.	Here isolation diffusion is required hence historically it is one generation behind as far as packing density is concerned
This is slower. The best clocking speed is 3GHz	BJT is historically much faster . The state of art in BJT is achieving 300GHz transit frequency.
Noise Figure is much better. This has no SHOT	Noise Figure is poorer because of SHOT noise,

noise , no PARTITION noise. It has flicker noise. For Low Noise Amplifier cryogenically cooled MESFET.	PARTITION noise and THERMAL noise.

I_D(μA)	g_m(×10^{-3} S)	r_{ds}(MΩ)	μ	I_C(μA)	g_m(×10^{-3} S)	r_π(kΩ)	r_0(MΩ)	μ
1	0.0478	85.2	4060	1	0.04	2500	85	3400
10	0.151	8.52	1280	10	0.4	250	8.5	3400
100	0.476	0.852	406	100	4	25	0.85	3400
1000	1.51	0.085	128	1000	40	2.5	0.085	3400
10,000	4.76	0.0085	40	10,000	400	0.25	0.0085	3400

HIGH FREQUENCY MODEL OF FET

In Figure 4, high frequency model of MOSFET is set up for calculating the short circuit current. We follow the same procedure as we followed in CE BJT.

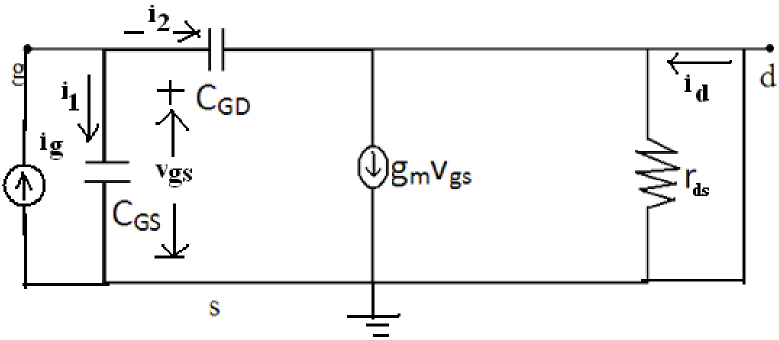


Figure4 . The incremental model of MOSFET at high frequency

There are various values for C_{gs} and C_{gd} in Triode Region, in Pentode Region and in Cut-off Region. In Analog Electronics for linear applications we are concerned only with Pentode Region. So we will consider the two capacitances in Pentode or in Saturation Region.

$C_{GS} = (2/3)\, C_{GC} + C_{GSO}.W$

$C_{GD} = C_{GDO}.W$

Where $C_{GC} = C_{OX}.WL$ and $C_{OX} = (\varepsilon_0 .\ \varepsilon_{oxide})/D_{oxide}$ F/cm2;

C_{GSO} = overlap capacitance between Gate and Source;

C_{GDO} = overlap capacitance between Gate and Drain;

Since r_{ds} is shorted therefore :

$I_2(j\omega) + I_d(j\omega) = g_m . V_{gs}(j\omega)$ 3

$V_{gs}(j\omega) = I_g(j\omega)/(j\omega(C_{GS} + C_{GD})$ 4

From Eq3 and Eq4:

$$\beta(j\omega) = \frac{I_d(j\omega)}{I_g(j\omega)} = \frac{g_m(1-\frac{j\omega C_{GD}}{g_m})}{j\omega(C_{GS}+C_{GD})}$$

5

Eq 5 can be rewritten as :

$$\beta(j\omega) = \frac{\omega_T}{j\omega}\left(1-\frac{j\omega}{\omega^*}\right)$$

6

Where $\omega_T = g_m/(C_{GS} + C_{GD})$ = Unity Gain BW or Gain × BW or Gain BW Product(GBP);

And $\omega^* = g_m/C_{GD}$ and $\beta(j\omega)$ falls at 20dB per decade.

Through proper substitutions in GBP expression we get :

$$\omega_T = \frac{g_m}{(C_{GS}+C_{CD})}$$

$$f_T = \frac{3}{2}\mu_n \frac{(V_{GS} - V_{TN})}{L^2}$$

A reduction in channel length by a factor of 10 increases f_T by 100.

In BJT GBP $< 1/(r_x C_\mu)$

BODY EFFECT or SUBSTRATE SENSITIVITY

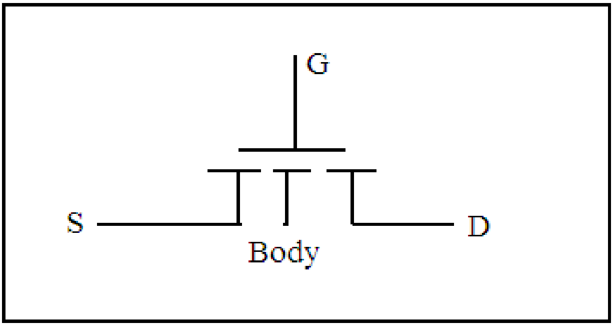


Figure 5. The symbol of (E)NMOS.

note there are 4 terminals: three for Source, Gate and Drain and fourth for the substrate or the body.

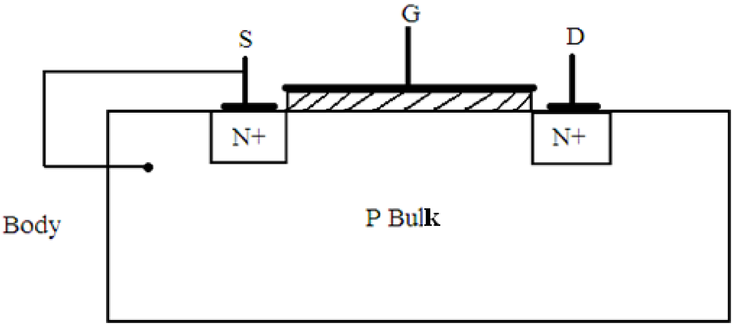


Figure 6. NMOS with source tied with body.

$$When\ V_{SB} = 0V, V_{TN} = 1V$$

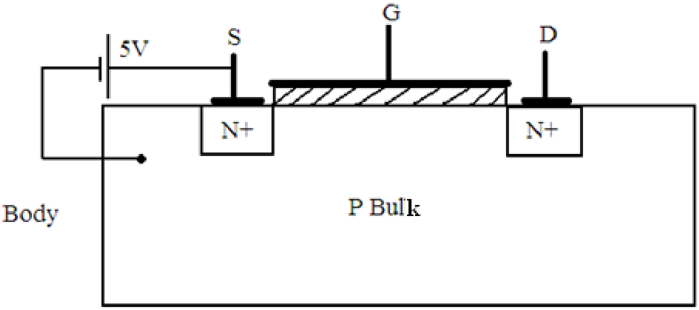


Figure 7. Here there is Source to Substrate bias of 5V.

$$When\ V_{SB} = 5V, V_{TN} = 2V$$

$$V_{TN} = V_{TO} + \gamma\left(\sqrt{V_{SB} + 2\Phi_F} - \sqrt{2\Phi_F}\right)$$

Where

$$V_{TO} = Zero\ substrate\ bias\ value\ for\ V_{TN}(V)$$

$$\gamma$$

$= Body\ Effect\ parameter$

$$(\sqrt{V})$$

$$2\Phi_F = Surface\ Potential\ Parameter$$

(V)

Typical Values are :-

$$-5V \leq V_{TO} \leq 5V$$

$$0 \leq \gamma \leq$$

3

$$\sqrt{V}$$

$$0.3V \leq 2\Phi_F \leq 1V$$

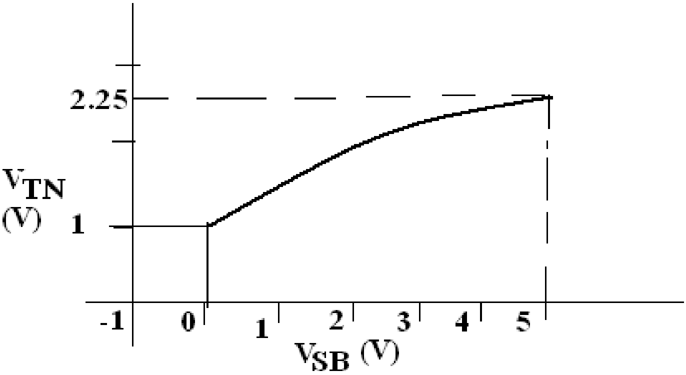


Figure 8. Threshold Voltage vs Source to Bulk voltage.

Chapter 4_Section 4.4._Design Rules for Fabricating nJFET.

Chapter 4_Section 4.4_gives the design procedure for JFET.

Chapter 4_Section 4.4_Design Rules for nJFET.

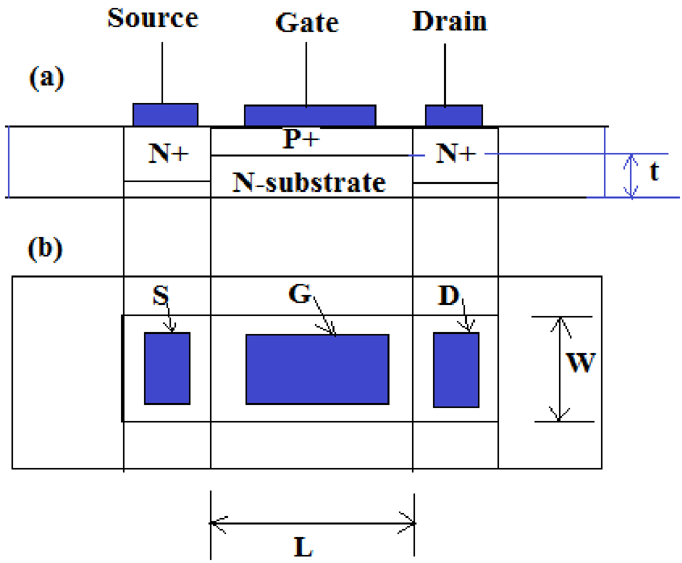


Figure 4.4.1. (a) Cross-sectional view of nJFET (b)Top view of nJFET.

We have already discussed the structure and physics of nJFET in m38382.

We saw that in Triode Region (Ohmic Region):

$$i_D = K_n[2(V_{GS} - V_P)v_{DS} - v_{DS}^2] \qquad for\ 0 < v_{DS} < (V_{GS} - V_P) \qquad 4.4.1.$$

V_P = Pinch – Off voltage of the channel given by the following expression:

$$V_p = \left(\frac{qN_D}{2\epsilon}\right)t^2 - \emptyset_{BO} \tag*{4.4.2}$$

For small drain-source voltages ($v_{DS} < 0.1V$):

$$i_D = K_n[2(V_{GS} - V_P)v_{DS}] \tag*{4.4.3.}$$

Therefore N-type Channel of nJFET offer linear resistance given by the following expression:

$$\frac{1}{R_{channel}} = \frac{i_D}{v_{DS}} = K_n[2(V_{GS} - V_p)] \tag*{4.4.4.}$$

Thus at small voltages Gate Controlled linear resistances can be realized using a JFET.

In Pentode Region, the drain current saturates at $V_{DS^*} = (V_{GS} - V_P)$:

$$i_{DS} = K_n(V_{GS} - V_P)^2 \tag*{4.4.5.}$$

At $V_{GS} = 0$, $i_{DS} = I_{DSS}$ and from Equation 4.4.5:

$$I_{DSS} = K_n(-V_P)^2 = \frac{V_P}{R_0/3} \quad where\ R_0 = \frac{\rho L}{W(t - d_0)} \tag*{4.4.6}$$

Here d_n = depletion penetration in N-substrate at zero gate voltage and is given by the following expression:

$$d_0 \approx d_n = \sqrt{\frac{2\epsilon_{Si}}{q} \times \frac{1}{N_D} \times \emptyset_{BO}} \qquad 4.4.7.$$

The customer will specify the pinch-off voltage and I_{DSS}. In practice these can range from

0 to -25V for Pinch-Off voltage and I_{DSS} = 10μA to 10A.

From these specifications the fabrication parameters and geometrical dimensions can be determined.

Section 4.4.1.Design nJFET for pinch-off Voltage -1V and -2V.

Keep the Gate P+ doping at 10^{19}/cc. Keep L=10μm and W=10μm.

Calculate the total thickness 't' for pinch-off voltage at -1V and -2V. Also calculate the penetration of depletion width in the channel at Gate Voltage = 0V. This Data will be required while calculating I_{DSS}.

Design procedure for V_P = -2V.

Using Eq.4.4.2. for N_D = 10^{13}/cc we get t = 18.4μm and using Equation 4.4.7 we get d_0=9.47μm.

At N_D = 10^{13}/cc the resistivity of the channel is 430.45Ω-cm. Using Equation 4.4.6. we get :

I_{DSS} = 12.45μA.

If instead of N_D = 10^{13}/cc , N_D = 10^{14}/cc is used then we get t = 5.9μm and using Equation 4.4.7 we get d_0 = 3 μm.

At N_D = 10^{14}/cc the resistivity of the channel is 43.45Ω-cm. Using Equation 4.4.6. we get :

I_{DSS} = 40μA.

Design procedure for V_P = -1V.

Using Eq.4.4.2. for N_D = 10^{13}/cc we get t = 14.6μm and using Equation 4.4.7 we get d_0 = 9.47μm.

At N_D = 10^{13}/cc the resistivity of the channel is 430.45Ω-cm. Using Equation 4.4.6. we get :

I_{DSS} = 3.47μA.

If instead of N_D = 10^{13}/cc , N_D = 10^{14}/cc is used then we get t = 4.7μm and using Equation 4.4.7 we get d_0 = 3 μm.

At N_D = 10^{14}/cc the resistivity of the channel is 43.45Ω-cm. Using Equation 4.4.6. we get :

I_{DSS} = 12μA.

SSPD_Chapter 5_Section 5.1. Photo-Diodes and its Responsivity Curves.
Chapter 5 is on Photonics which will cover all the optical sources as well as optical sensors.Chapter 5,Section 5.1, covers the Photo-diode which can act as optical sensor under reverse bias condition and acts as the photo-voltaic cell under no bias condition.

Chapter 5_Section 5.1_Semi-conductor Materials and its Absorption Coefficient curves.

In Photonics we study optical energy sources and optical energy detectors/sensors. Both are diodes but they are especially constructed so that a source may emit a copious amount of optical energy and detector may absorb a 99% of the incident optical energy to produce a optimum response. The optical energy detectors are always kept reverse biased but when there is no bias then under illumination condition they can be used as photo-voltaic cells or solar cells also. The optical detectors under reverse bias detects a particular wavelength or shorter of Electro-Magnetic Radiation only when the Energy Packet of the given EM Radiation contains (hν) energy greater than the band-gap of the semiconducting material constituting the optical detectors.

There is a very convenient relationship which directly gives the threshold wavelength . Incident light shorter than this threshold wavelength will optically excite optical sensor and generate an electrical signal in response to photo-excitation. The relationship is as follows:

$$\lambda_{Thrshold}(\mu m) = \frac{1.24}{E_G(eV)} \qquad 5.1.$$

The sensitivity of the Optical Detector depends on the absorption coefficient of the material constituting the optical detector. The absorption coefficient (α per cm) is described in Figure 5.1. as formulated below:

$$I_t(transmitted\ light) = I_0(incident\ light)Exp[-\alpha.t] \quad where\ \ \alpha\ is\ the\ absorption\ coeeficient \qquad 5.2.$$

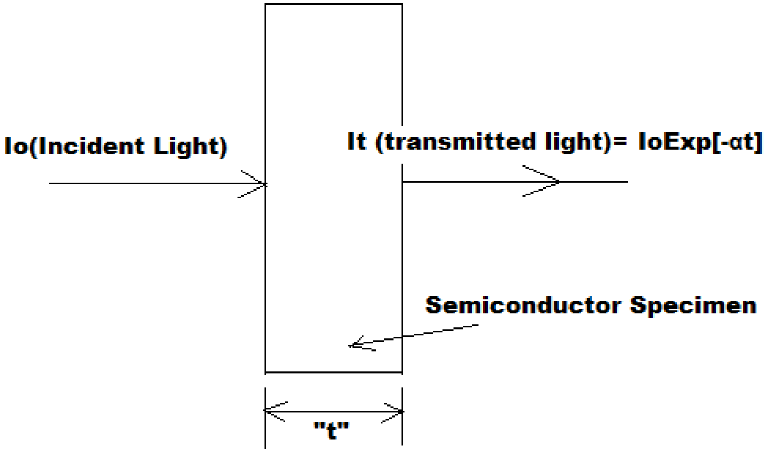
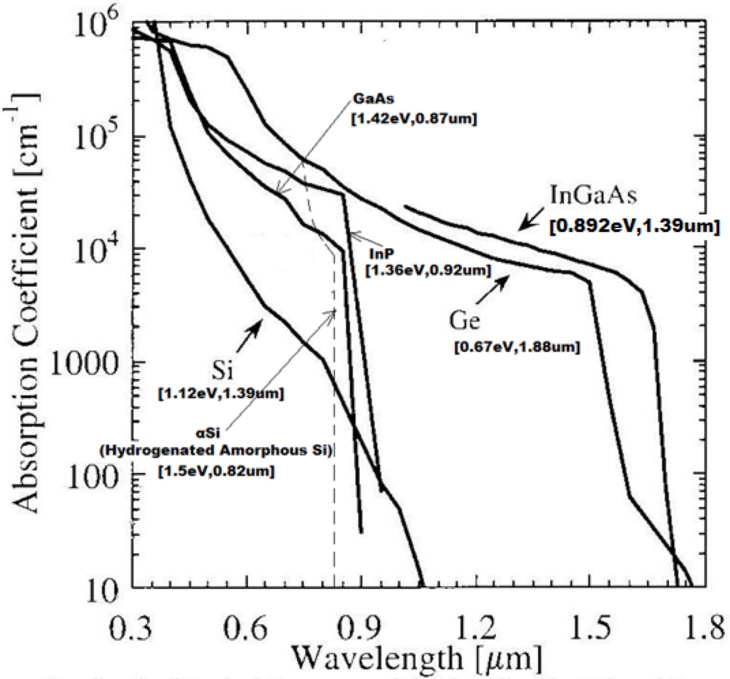


Figure 5.1. Schematic Diagram of the incident light and emerging **light from a semiconductor specimen of thickness "t".**



Handbook of Optical Constants of Solids, edited by Edward D. Palik, (1985), Academic Press NY.

Figure 5.2 Absorption Coefficient Curves of Different Semi-conductor Materials.
[The Band Gap and the threshold wavelength are given in the square bracket next to the chemical symbol of the semiconductor material]

The absorption coefficient curves are shown in Figure 5.2. for elemental semi-conductors such as Ge, Si and Hydrogenated Amorphous Si (α-Si) and for compound semiconductors such as GaAS, InP and InGaAs.

As is evident from Figure 5.2, compound semiconductors have a very sharp cut-off in absorption coefficient at wavelengths larger than threshold wavelength.

GaAs has a Band-Gap of 1.42eV which gives a threshold wavelength of 0.87μm from Equation 5.1. Hence for wavelengths longer than 0.87μm, GaAs is transparent. Light passes through a specimen of GaAs unobstructed at wavelengths longer than 0.87μm. At 0.87μm the specimen becomes opaque and absorbs light at 0.87μm and shorter wavelengths.

The reciprocal of the absorption coefficients is penetration depth as defined below:

$$penetration\ depth = d^* = \frac{1}{\alpha} = \frac{1}{10^4} cm = 1\mu m\ for\ GaAs \qquad 5.3.$$

Penetration Depth tells that the thickness of GaAs detector/sensor must be greater than d* = 1μm in order to fully absorb the incident photons and give a large electrical response. All compound semiconductors behave in identical manner. As is evident from Figure 5.2, GaAs, InP and InGaAs give a step jump in the absorption coefficient at and shorter wavelength than threshold wavelengths namely at 0.82μm, 0.92μm and 1.39μm respectively. For optimum electrical response the thickness of absorbing layer must be about 1μm in all three cases. Compound Semiconductor exhibit such a sharp response due to the fact that they are Direct-Bandgap Semi-conductor.

In contrast elemental semi-conductors such as Si and Ge have much softer response. In case of Si, at wavelength 1.39μm and at shorter wavelengths the response only gradually increases as a result maximum response can be obtained only if the thickness of active absorbing layer is 50μm. In case of Ge, response gradually increases at 1.88μm and at shorter wavelengths and here also the effective thickness should be of the order of 50μm. This increases the cost of the detector or photo-voltaic cell.

In Si, absorption coefficient can be vastly improved if we use hydrogenated amorphous Si(α-Si). This has an identical responsivity curve as that of GaAs hence it is the commonly used material in Solar Cells.

Chapter 5_Section 5.2. Direct and in-direct Band Gap

Chapter 5_Section 5.2 describes the difference between Direct and In-direct Band-Gap Materials and shows that Direct Band Gap are suitable for Optical Sources whereas bot Direct and In-direct can be used for Optical Detectors.

Chapter 5_Section 5.2. Direct and in-direct Band Gap Materials.

Section 5.2.

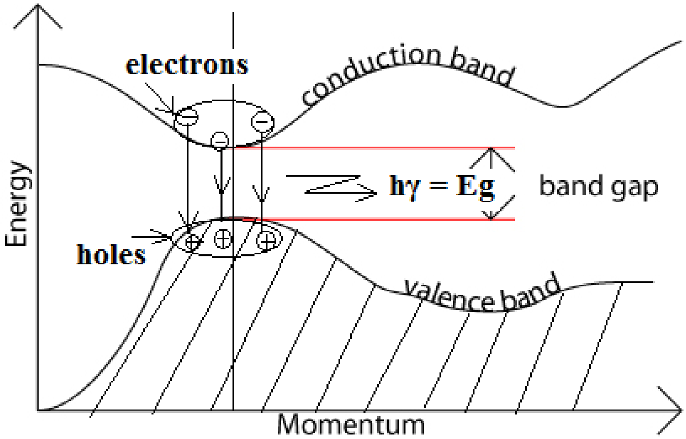


Figure 5.3. Direct recombination of electrons and holes mediated by photons of energy packet = Eg=hγ.

As seen in Figure 5.3, in Direct Band Gap materials the Energy Minima of Conduction Band and Energy Maxima of Valence band lie along the vertical line which passes through identical momentum value. So when electron in conduction band and hole in valence band recombine then there is : $\Delta E = E_g$ and $\Delta p = 0$. Therefore a photon which has an energy packet $= h\gamma = E_g$ and momentum 'p' = negligible, such a photon can successfully mediate such a recombination. These are called Radiative Transitions where for every electron and hole recombination there is the emission of a photon.

Therefore we say that Vertical Transitions cause Radiative Transitions or Emissive Transitions. In such Direct Band-Gap materials, internal Quantum Efficiency is 100% and if there are 1000 EHPs recombining then through Vertical Transitions 1000 photons of requisite wavelength given by Equation 5.1 are emitted.

In Figure 5.3., E-p or E-k diagram of a In-direct Band Gap material is given.AS can be seen the electrons at the bottom of conduction band and holes at the top of the valence band must recombine obliquely. They cannot recombine vertically. This oblique recombination is called in-direct recombination. In this oblique recombination ,for the conservation law to hold good after the recombination of electrons and holes, extra Energy $\Delta E = E_g$ and extra $\Delta p = p2-p1$, both have to be carried away. Photon has very little momentum. So only Phonons can do this job. It can take away the extra Energy as well as the extra Momentum generated by in-direct recombination.

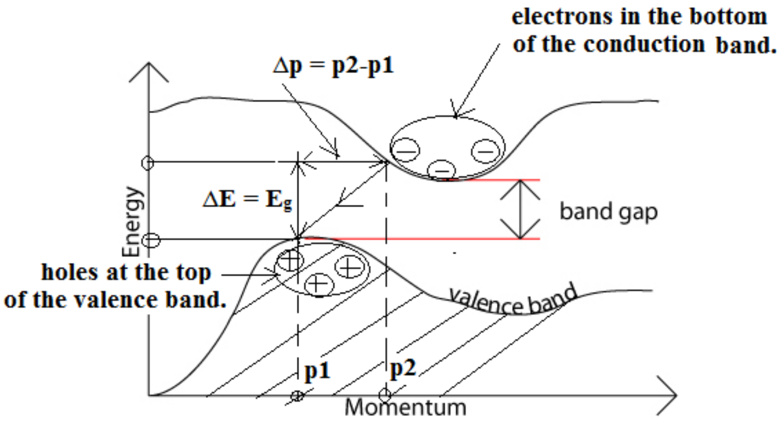


Figure 5.4. The E-p or E-k diagram of Indirect Band Gap Semiconductors.

Now the question arises as to what is phonons.

Phonons are quanta of lattice vibrational energy. They travel with the velocity of sound in the dielectric medium. The phonons which travel as longitudinal waves are called acoustic mode phonon and phonons which travel as transverse waves are called optical mode phonon. Here I may just refresh the terms “longitudinal” and “transverse”. Light travel as a Transverse Wave. Here the medium moves transverse to the direction of propagation. Whereas Sound transmission medium moves in the same direction as the direction of propagation. Here I may also point out that in stable state every component of a system is in minimum energy state. Electrons in the bottom of Conduction Band and holes at the top of the Valence Band are also in minimum energy state.

Since all elemental semi-conductors have in-direct Energy Band Diagram, hence their EHPs recombination or transitions are always accompanied with increased lattice vibration which is tantamount to heating. Hence here we have non-radiative transition and Quantum Efficiency is zero.

Since Compound Semiconductors are Direct Band-Gap materials they are suitable as Optical Sources. But Optical Detectors can be either Direct or In-Direct Band Gap material since both respond to optical excitation except that in-Direct material will have to be thicker to be able to give equivalent response for a given optical input.

SSPD_Chapter 5_Section 5.3. Spectal Responsivity of Photo-Diodes.
SSPD_Chapter 5_Section 5.3.and Section 5.4. describe the parameter Spectral Responsivity and its optimal structure for maximum responsivity.

SSPD_Chapter 5_Section 5.3. Spectal Responsivity of Photo-Diodes.

Spectral Responsivity is the main parameter which decides the application of Photo-diodes of given Spectral Responsivity curves. Here we analyze the theoretical formulation of this parameter and then we study the design of Photo-Diodes for various applications.

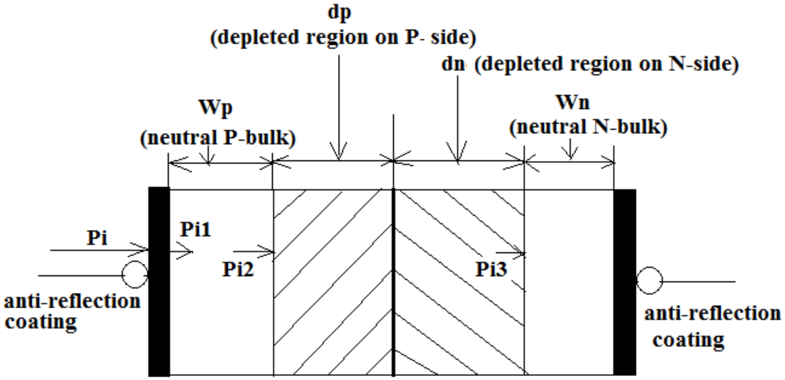


Figure 5.3.1. Cross-sectional view of PN-Junction Photo-Diode. Shaded region is the depletion region which is the active region.Photo-generated EHPs in this depleted region contribute to the photo-current.

Electron-Hole Pairs (EHPs) generated in the neutral bulk travel by diffusion whereas EHPs generated in depleted region travel by drift. As shown below diffusion velocity is two orders of magnitude smaller than the drift velocity:

$$Drift\ Velocity = V_{drift} = \mu_n E = (1400\frac{cm^2}{V-s}) \times 1000\frac{V}{cm}) = 1.4 \times 10^6 \frac{cm}{s} \tag{5.3.1.}$$

$$J_{diffusion} = -qD_n\frac{\partial n}{\partial z} = qD_n\frac{\hat{n_0}}{L_n} = qV_{diff}\frac{\hat{n_0}}{2}; \tag{5.3.2.}$$

Therefore:

$$V_{diff} = \frac{2D_n}{L_n} = \frac{2 \times 36.4cm^2/s}{50 \times 10^{-4}cm} = 1.456 \times 10^4\frac{cm}{s} \tag{5.3.3.}$$

Hence bulk generated EHP slow down the photo-response. Therefore contribution of bulk-generated EHPs to photo-current is kept at the minimum. The depleted region of the reverse biased diode is kept at the maximum and is kept as the active contributor of the photo-current.

For theoretical formulation of the Spectral Responsivity let us examine Figure 5.3.1.

$$The\ power\ absorbed\ by\ the\ Active\ Region = P_{i2} - P_{i3} = P_{abs}(\frac{J}{s}) \tag{5.3.4.}$$

The intervening Bulk Region always brings down the effective absorption of the incident light. To ensure that all the incident quanta get absorbed, we use anti-reflection coating on the front surface of the photo-diode and we minimize the neutral bulk region widths.

Let internal quantum efficiency = η (no. of EHPs generated per photon).

Ideally 100photons should generate 100EHPs and this implies $\eta = 1$ but practically $\eta < 1$.

$$No.\ of\ absorbed\ photons\ per\ second\ in\ the\ active\ region = \frac{P_{abs}}{h\nu}\frac{photons}{s} \qquad 5.3.5.$$

$$Photo - generated\frac{EHPs}{s} = \frac{\eta P_{abs}}{h\nu}\frac{EHPs}{s} \qquad 5.3.6.$$

Each EHP gives rise to ‘q’ coulombs of circuital charge hence effective photo-generated current is:

$$I_{photo} = q \times \frac{\eta P_{abs}}{h\nu}\frac{Coulombs}{s} \qquad 5.3.7.$$

Now spectral Responsivity is defined as:

$$R = \frac{I_{photo}}{P_{abs}} = \frac{q\eta}{h\nu} = \frac{q\eta}{hc} \times \lambda = 8065.545\eta\left(\frac{A}{W - cm}\right) \times \lambda(cm) \qquad 5.3.8.$$

In Equation 5.3.8. there is no threshold term as predicted by Equation 5.1. Hence Equation 5.3.8. gives the dotted curve in Figure 5.3.2. which is an idealized Spectral Responsivity curve. The real Responsivity Curve is given by Bold Curves. Table 5.3.1. tabulates the coordinates of the idealized curves for different photo-diodes[($\lambda_{\text{Threshold}} = 1000$nm, $\eta = 0.9$), ($\lambda_{\text{Threshold}} = 1500$nm, $\eta = 0.5$), ($\lambda_{\text{Threshold}} = 1700$nm, $\eta = 0.7$),]

Table 5.3.1. Responsivity data(A/W) for 3 Photo-Diodes of given $\lambda_{\text{Threshold}}$ & η.

λ(cm)	500×10^{-7}	600×10^{-7}	700×10^{-7}	900×10^{-7}	1000×10^{-7}	1500×10^{-7}	1900×10^{-7}	$\lambda_{\text{Threshold}}$
R($\eta=0.9$)	0.36	0.435	0.508	0.653	0.725	1.09	1.38	1000nm
R($\eta=0.5$)	0.202	0.242	0.282	0.363	0.403	0.605	0.766	1500nm
R($\eta=0.7$)	0.282	0.339	0.395	0.508	0.564	0.846	1.073	1700nm

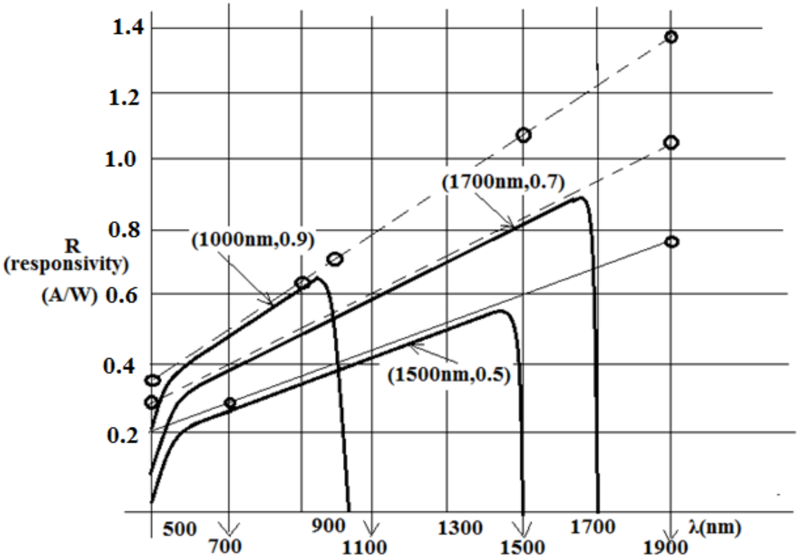


Figure 5.3.2. Spectral Responsivity Curves (Idealized and Real) for photo-diodes with (λthreshold=1000nm, η=0.9), (λthreshold=1700nm, η=0.7) and (λthreshold=1500nm, η=0.5).

Experimental curves for a given photo-diode ($\lambda_{\text{Threshold}}=1000$nm, $\eta=0.8$) are given in Figure

5.3.3. over a spectral range of 400nm to 1100nm(blue curve) and over a range of 200nm to 1100nm(red curve). As can be seen in Figure 5.3.2. as well as in Figure 5.3.3. the Spectral Responsivity Curve sharply dips in the region of threshold wavelength($\lambda_{\text{Thrshold}}$) given by Equation 5.1. as expected. At longer wavelengths the photons are not energetic enough to cause photo-excitation.

In Figure 5.3.2. at shorter wavelengths also there is departure from the idealized curve. At shorter wavelengths , absorption coefficient (α) sharply increases as seen in Figure 5.2. Hence penetration depth becomes shallow which leads to reduced absorbed power resulting in reduced response. But in Figure 5.3.3. in UV region the curve experiences a bump instead of a dip. This point s not clear to me at this point.

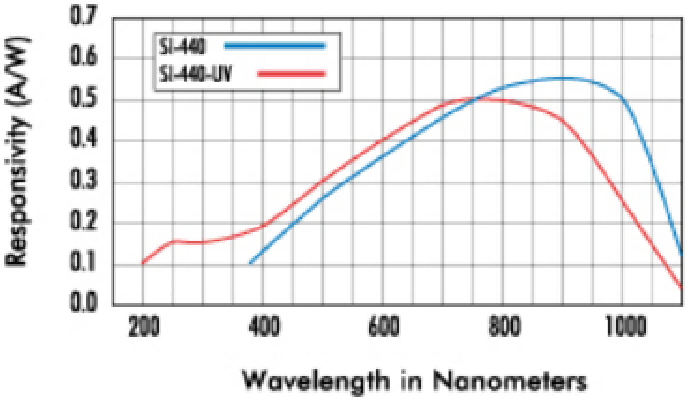


Figure 5.3.3. The Spectral Responsivity Curve for a given diode over a range of 400nm to 1100nm (blue) and over a range of 200nm to 1100nm(red).

5.4. Design of a Photo-Diode in different parts of the Electro Magnetic Spectrum depending on the application.

Different applications require maximal response in different parts of SPECTRUM. For instance Night vision requires maximum response in Near IR(800nm to 1000nm).

1. Metal-Semiconductor photo-diodes have best internal quantum efficiency and maximal response in UV and Visible part of the spectrum. The Metal-Semiconductor photo-diodes available are Au-Si, AgZn-Si.
2. Si photo-diodes (with anti-reflection coating) have best internal quantum efficiency and maximal response in Near IR part of the spectrum(800nm to 900nm).
3. Ge photo-diodes and GaInAs have best internal quantum efficiency and maximal response in IR part of the spectrum(1000nm to 1600nm).
4. Cryogenically cooled InSb and InS photo-diodes at 77K have best internal quantum efficiency and maximal response in 2 to 8 microns wavelength.

5.4.1. PIN Diode.

For maximal response of the photo-diodes in faintest of light, various structural modifications are adopted. The first modification is the introduction of intrinsic region between P and N region. As seen in Figure 5.4.1. a PIN diode has a depletion region in the entire Intrinsic Region. This gives a direct control over the depletion region and the active region can be tailored according to our requirements.

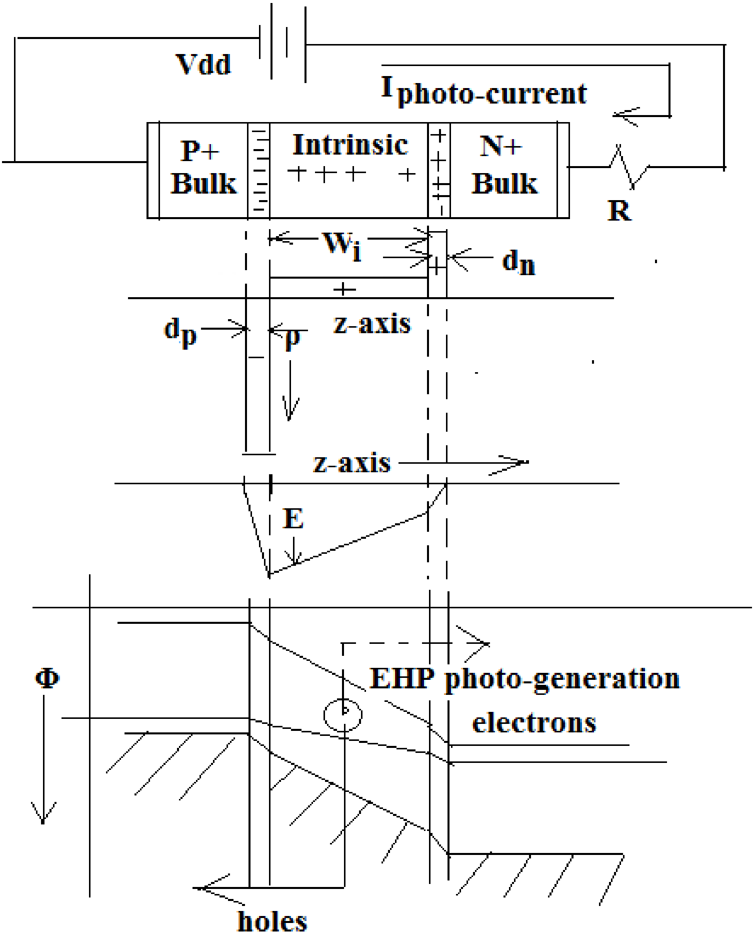


Figure 5.4.1. A PIN Diode reverse biased, Space-Charge distribution, Field Distribution, Energy-Band Diagram and EHP photo-generation and its contribution to the reverse phot-current.

As seen in Figure 5.4.1. the active layer comprises of $W_{active} = d_p + W_i + d_n$ .This increase in active layer by the addition of intrinsic layer increases the responsivity mani-fold. Also the large drift field through out the intrinsic region is effective in quickly removing the photo-generated EHPs. This adds to the speed of the response of the photodiode.

5.4.2. Avalanche Photo-diode (APD) or Reach Through Avalanche Photo-diode.

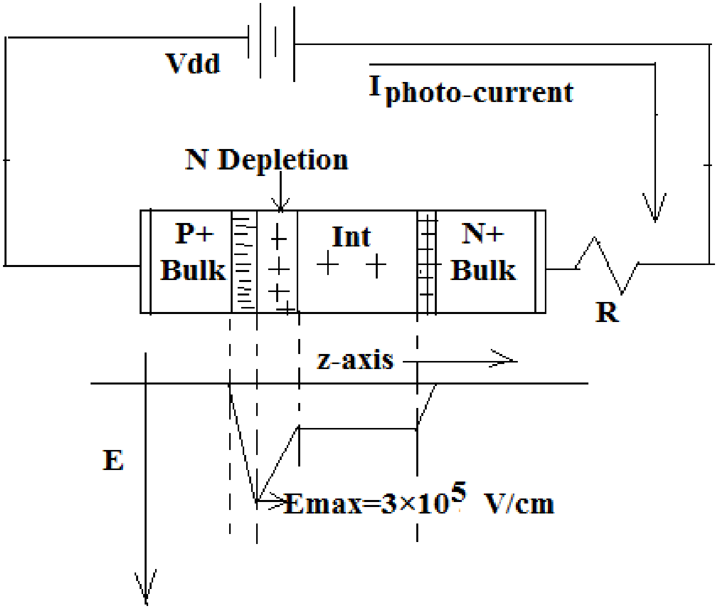


Figure 5.4.2. Avalanche Photo-diode. By introducing a lightly doped N-layer between P+ and Intrinsic region the Emax reaches the critical magnitude where Avalanche Multiplication Factor is introduced. This causes EHP multiplication and effectively acts as a photo-multiplier tube.

As seen in Figure 5.4.2., a fourth layer of lightly doped N-layer considerable increases the Emax and it crosses the critical field and introduces Avalanche Multiplication Factor (M). This effectively multiplies the photo-current as it did in photo-multiplier tube in Vacuum Tube Era but it introduces considerable amount of noise due to the statistical nature of Avalanche Multiplication. This Noise can be suppressed by making the ionization rate due to Electron and Hole vastly unequal.

In a Si-APD it is possible to make electron ionization rate = 100 times hole ionization rate. This considerable suppresses the Multiplication Noise.

In Ge, GaAs and in InGaAS we have electron ionization rate = 10 times hole ionization rate hence here there is considerable noise and hence it cannot operate in very faint light environment.

The Photo-current multiplication graph is shown in Figure 5.4.3.

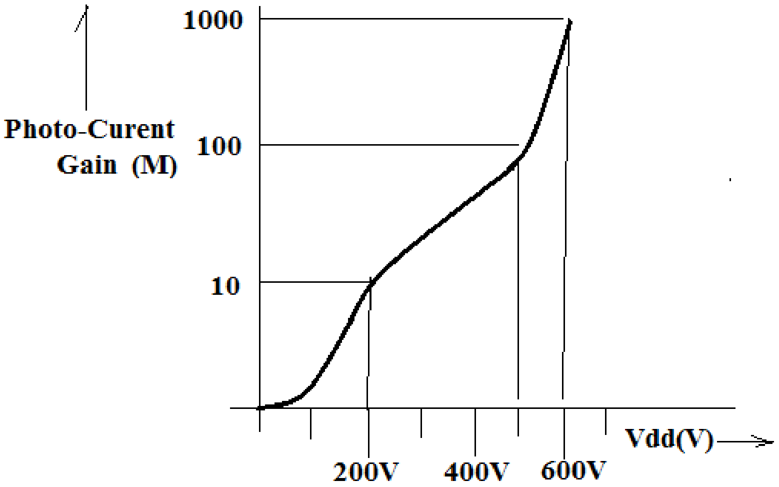


Figure 5.4.3. Photo-Current Gain due Multiplication Factor with increase in applied reverse bias across the APD.

M=1/[1-(Vdd/BVdo)]^n
Vdd= applied reverse bias
BVdo= Avalanche Break-down Voltage
M=Avalanche Multiplication Factor
n = Miller Indices= 2 to 5

SSPD_Chapter 5_Section 5.3_Supplement_Photo-Diode_I-V curve.
SSPD_Chapter 5_Section 5.3. Supplement gives the I-V family of curves of Photo-Diode under dark and illuminance condition.

SSPD_Chapter 5_Section 5.3_Supplement_Photo-Diode_I-V curve.

The I-V Characteristics of a Photo-Diode is exactly the same as a classical diode except that there is an added family of curves due to illumination. Under reverse-bias, there is the DARK classical reverse leakage current which is of ‘nA’ magnitude. Under illumination condition, there is a family of curve under different illuminance condition as predicted by Equation 5.3.7 :

$$I_{photo} = q \times \frac{\eta P_{abs}}{h\nu}\,\frac{Coulombs}{s} = 8065.545\eta(\frac{A}{W-cm}) \times P_{abs}(W) \times \lambda(cm) \qquad 5.3.7.$$

Using Equation 5.3.7. we generate a Table 5.3.2. for a photo-diode which has internal quantum efficiency = 0.9 and it is being excited by 1000nm IR light under different illuminance condition.

Table 5.3.2. Generation of Photo Current in reverse-biased Photo-Diode with internal quantum efficiency = 0.9 and irradiated by 1000nm IR light under different illuminance condition.

Pabs (μW)	Iphoto(μA)
10	7.259
20	14.5
100	72.6

Assuming that we are using Ge-Photo Diode, the family of I-V curves under dark condition and under various illuminance condition is given in Figure 5.3.4. As seen in the Figure 5.3.4. IIIrd Quadrent is Photo-Conductive Mode, IVth Quadrent is Photo-Voltaic Mode and Ist Quadrent is Photo-Emissive Mode only for Compound Semi-Conductors.

In Photo-Conductive Mode, under reverse bias condition Photo-Diode conducts and acts as a Constant Current Source(CCS) to different extents according to the Illuminance condition. Under Dark Condition, CCS is open. With 10μW illuminance it acts as CCS of 7.259μA, with 20μW illuminance it acts as CCS of 14.5μA and with 100μW illuminance it acts as CCS of 72.59μA.

In fourth quadrent, Photo-Diode has no Bias Voltage. Under 100μW illuminance, the Y-intercept of the corresponding I-V curve gives the short-circuit current (ISC) generated by the photo-cell and the X-intercept of the same curve gives the open-circuit voltage (VOC) generated at the open circuit terminals of the photo-cell. When the Photo-Cell is terminated by load resistance RL then there is load line of slope = Tan θ = 1/RL. . This Load Line intersects the I-V curve of the photo cell at Q(I_L,VI_L). This Q-point tells us the current delivered to the terminated load. There is a particular Q-point on the I-V curve which maximizes the output power (I × V). A load matching circuit ensures that the cell operates at that point.

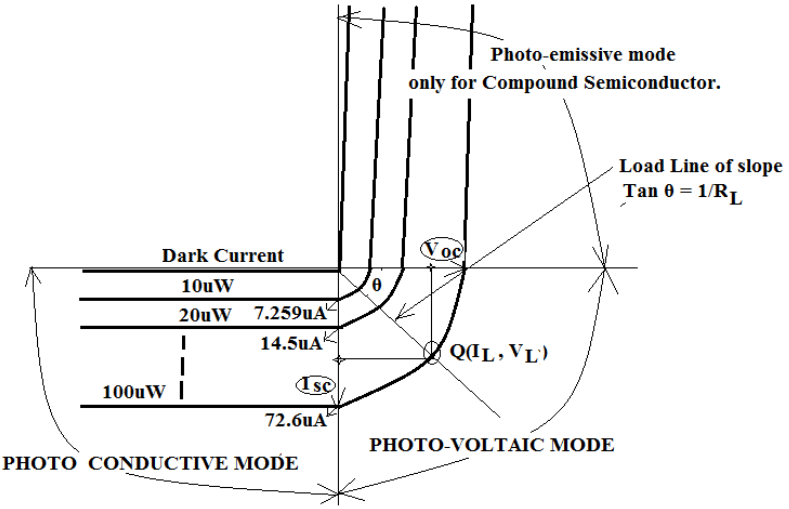


Figure 5.3.4. I-V Family of Curves of a Photo-Diode under diffrent Illuminance condition.

“Solar Energy is diffuse. Each square meter of solar cells can produce 25W of continuous power when averaged over day, night and sunny/cloudy days. The average electricity consumption of the world is over 10^{12}W. If all he electricity is to be produced by solar cells, the cells need to cover a huge 200km by 200km area. Therefore low cost and high energy conversion efficiency, defined as the ratio of the electric output to the solar energy input, are important. Solar cells can be made of amorphous or polycrystalline as well as single-crystalline semiconductors. The first two types are less expensive but also less efficient in electricity generation”.[“Modern Semiconductor Devices for Integrated Circuits”, by Chenming Calvin Hu, Pearson 2010].

“The best solar cell efficiency(~24%) is obtained with E_g values between 1.2eV and 1.9eV. Commercial rooftop silicon solar cell panels have conversion efficiencies between 15 to 29%. Tandem solar cells can achieve very high (over 30%) energy conversion efficiency by using two or more semiconductor materials. One with a large E_g absorbs and converts the UV portion of the solar insolation to electricity and another of smaller E_g material, positioned behind the first, does the same to the solar insolation that is not absorbed by the first material.”

$$Fill\ Factor(FF) = \frac{(I \times V)_{Max}}{I_{SC} \times V_{OC}} \qquad 5.3.9.$$

FF is typically 0.75. The short-circuit current, I_{SC}, is the photo-current given in Equation 5.3.7. and is directly proportional to the Illuminance condition.

Section 5.3.1. Application of Photo-Diode as demodulator in Optical Fiber Communication.

Optical Fiber Communication(OFC) is rapidly supplanting the copper wired Communication because of inherent advantage of OFC. OFC offers almost an infinite Band-Width hence large number of channels can be Frequency-Divison-Multiplexed (FDM) or Time-Divison-Multiplexed (TDM). Plus it is very light weight, much less voluminous and hence very cheap in handling. Its long haul, heavy traffic property makes it very attractive proposition in terms of cheaply affordable communication means by common masses. Therefore Photonics has assumed a very big importance both as detectors as well as photo sources.

So the most popular application of Photo-Diode is as Demodulator in OFC as shown in Figure 5.3.5.

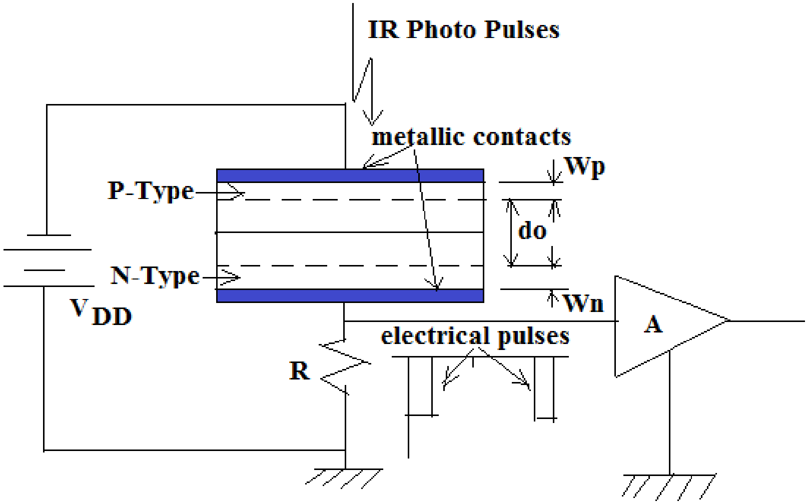


Figure 5.3.5. Photo-Diode Detector or demodulator circuit at the receiving end of the optical fiber communication system.

Photo-Pulses coming down the Optical Fiber Link are coupled into the photo-detector. In Photo-Diode, photo-pulses excite the photo-current pulses. These photo-current pulses, flowing through the resistance R of the detector circuit, produce voltage pulses. The voltage pulsesa are amplified by the Amplifier and passed on to Digital Signal Processing Chip.

SSPD_Chapter 5_Section5.5._Photo Sources.
SSPD_Chapter 5_Section5.5. describes the physics and design of LEDs of various colours.

SSPD_Chapter 5_Section5.5._Photo Sources.

As already discussed in Section 5.2. only compound semiconductors can act as photo-sources as they undergo Radiative Transition and Radiative Recombination.

Average time taken for radiative recombination = τ_{RR} ;

Average time taken for non-radiative recombination = τ_{NR} ;

Number of Radiative Recombinations /second = R_r = 1/ τ_{RR} ;

Number of non-Radiative Recombinations /second = R_{nr} = 1/ τ_{NR} ;

Internal Quantum Efficiency = η :

$$\eta = \frac{radiative\ recombinations/s}{total\ recombinations/s} = \frac{R_r}{R_r + R_{nr}} = \frac{\tau_{NR}}{\tau_{NR} + \tau_{RR}} \qquad 5.5.1.$$

As already discussed in Section 5.2., Compound Semiconductors with direct band-gap have high internal Quantum Efficiency whereas elemental semi-conductors have practically zero internal

Quantum Efficiency.

The principle of Photo-Sources (Light Emitting Diode as well as LASER Diode) is stated in three broad terms:

1. First the Compound Semiconductor has to be chosen of the correct band-gap so that the radiative transition generate the required wavelength as given by Equation 5.1.

$$\lambda_{Threshold}(\mu m) = \frac{1.24}{E_g(eV)} \qquad 5.1.$$

1. A Substrate of matching lattice parameter has to be chosen over which the epitaxial layer of the requisite material will be grown.
2. Third the Diode made of the required material must be driven hard in forward direction so as to create heavy injection condition. Under heavy injection condition there is copious amount of excess carriers available on both sides of the depletion layer. The life-time of the minority carriers should have short life-time so that they undergo rapid direct-recombination generating large amount of required wavelength photons. This process is shown in Figure 5.5.1.

Section 5.5.1. Design of 0.87μm (IR) Light-

Emitting Diode.

From Equation 5.1., for emitting 0.87μm the required Band –Gap is 1.42eV hence GaAs is the material of choice. In this case GaAs epitaxial layer is grown on GaAs Substrate thereby providing automatic lattice matching between the epitaxial layer and the substrate. We propose to fabricate Planar Surface Emitting LED hence we choose PN+ structure with P-layer being thinner than the reciprocal of absorption coefficient. PN+ structure ensures that majority of recombination takes part in P-Bulk and since P Bulk-Width is thinner than penetration depth hence most of the photons of $\lambda = 0.87\mu m$ manage to cross the P-Bulk and emerge out of the Planar Surface as shown in Figure 5.5.1. This is an optimum structure for Planar Surface Emitting LED. The near proximity of the emitting surface to the recombination region lowers the internal quantum efficiency because the surface states of the emitting surface leads to higher non-radiative recombinations.

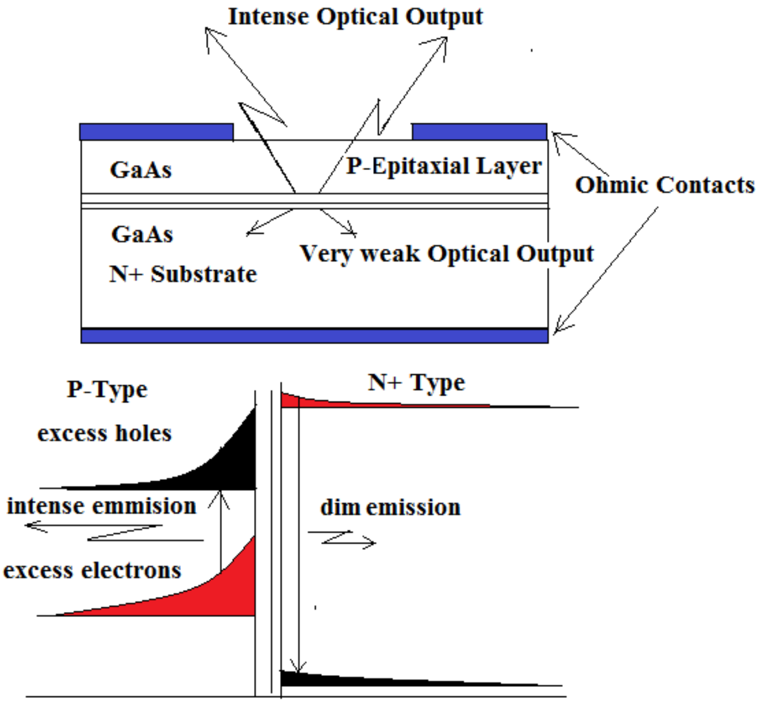


Figure 5.5.1.(a)Planar Surface Emitting LED structure.(b) Carrier Concentration Profile of PN+ Junction Diode.

This is a Homojunction LED since same material GaAs is used on two sides. This generally has a poor overall Quantum efficiency.

Section 5.5.2. Design of LED of various visible colours (Red, Orange,Yellow, Green and Blue).

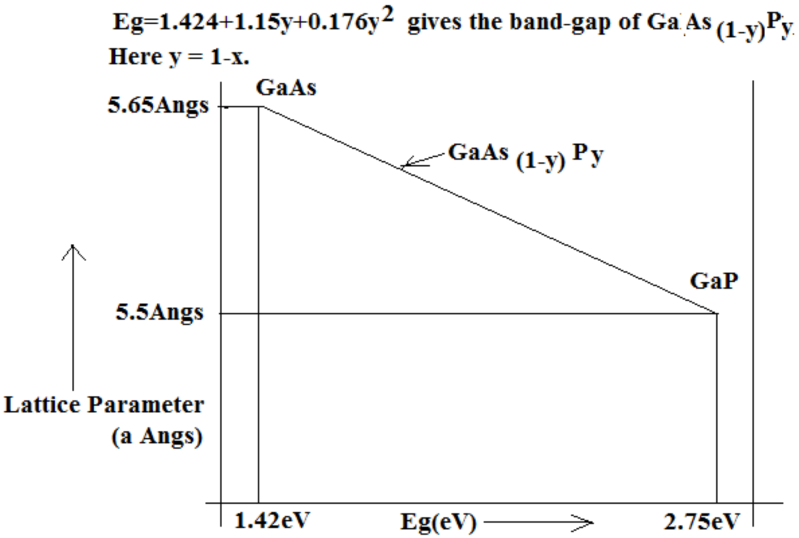


Figure 5.5.2.System Topology of Ternary Compound $GaAs_xP_{(1-x)}$ which is the constituent material of coloured LEDs.

The system topology of GaAsP ternary compound is given. Ga is Group III element and As and P are Group V element. As and P with stoichiometric coefficient (1-y) and y combine with Ga to form a tetravalent structure which is the basic crystalline structure of all semi-conductors. By controlling the stoichiometric coefficient ‘y’ the band-gap of the resulting ternary compound can be tailored by the relation given below:

$$E_g = (1.424 + 1.15y + 0.176y^2)eV \qquad 5.5.2.$$

Tailoring the band-gap by finely tuning the stoichiometric coefficient ‘y’ is known as *Band-Gap Engineering*. In the process of Band-Gap Engineering, the lattice parameter is also reduced from

5.65Angstrom for GaAs to 5.5Angstrom for GaP. We can use the relation ship (5.5.2) to determine the constituent materials and the substrate for fabricating LEDS of different colours namely Red, Orange, Yellow, Green and Blue. The results of thgis design is given in Table 5.5.1.

Red LED corresponds to 0.64μm which can be emitted from $GaAs_{(1-y)}P_y$ of $E_g = 1.9eV$. The substrate used is GaAs and stoichometric coefficient ‘y’ = 0.4.

Orange LED corresponds to 0.62μm which can be emitted from $GaAs_{(1-y)}P_y$ of $E_g = 2.0eV$. The substrate used is GaP and stoichometric coefficient ‘y’ = 0.5.

Yellow LED corresponds to 0.58μm which can be emitted from $GaAs_{(1-y)}P_y$ of $E_g = 2.14eV$. The substrate used is GaP and stoichometric coefficient ‘y’ = 0.58.

Green LED corresponds to 0.55μm which can be emitted from $GaAs_{(1-y)}P_y$ of $E_g = 2.35eV$. The substrate used is GaP and stoichometric coefficient ‘y’ = 0.656.

Blue LED corresponds to 0.475μm which can be emitted from $GaAs_{(1-y)}P_y$ of $E_g = 2.61eV$. The substrate used is SiC and stoichometric coefficient ‘y’ = 0.9.

Generally compound semi-conductors have direct band-gap and hence have high Internal Quantum Efficiency. But in ternary $GaAs_{(1-y)}P_y$ as ‘y’ exceeds 0.45, the ternary compound becomes In-direct and its Internal Quantum Efficiency drastically falls as shown in Figure 5.5.3.By doping this material with Nitrogen it is transformed into Direct Band-Gap material and internal Quantum efficiency is restored but it is not good enough for LASER Diodes.

By Nitrogen doping , traps are introduced vertically above the valence band peak in E-p or E-k diagram. These traps enable vertical transitions thereby improving internal Quantum Efficiency.

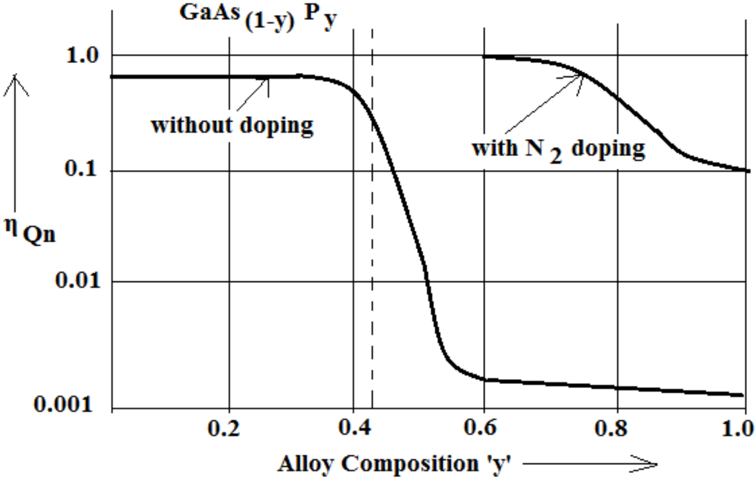


Figure 5.5.3. Internal Quantum Efficiency vs Stoichiometric Coefficient 'y'

Section 5.5.2.Homo-Junction, Hetreo-Junction and Double Hetero-junction.

Homojunctions are made of the same semiconductor material on both sides of the junction. Whereas Heterojunction are made of different band-gap material, but matching lattice parameters, on both sides of the junction. In Double Heterojunction one material of narrow bandgap is sandwitched between two other materials of wider bandgap. This leads to the formation of Quantum Well in the narrow bandgap material.

In a symmetric Homojunction , diode current is equally divided between electrons and holes and Injection Efficiency in any direction is 0.5. But in symmetric Heterojunction (symmetric in doping sense but highly asymmetric in Band Gap sense) , diode current is very unequally divided between electrons and holes as will become clear from Figure 5.5.4.

As seen in Figure 5.5.4. in Homojunction Diode with equal doping, the built-in Barrier potential seen by electrons on N-Side is the same as that seen by holes on P-Side. This is because $\Phi_{Bno} = \Phi_{Bpo}$. But in Heterojunction Diode (with very unequal Energy Band Gap) even with equal doping, $\Phi_{Bno} \neq \Phi_{Bpo}$. In Figure 5.5.4., Band Gap on P-Side (E_{g1}) is much larger than Band Gap on N-Side (E_{g2}). Hence $\Phi_{Bno} > \Phi_{Bpo}$. This means injection from P-Side is several orders of magnitude larger than the injection from N-Side. Hence Injection Efficiency is almost 100% from P-Side.

This becomes evident if we look at the Injection Efficiency Expression. From Equation 4.3.13:

$$\gamma = \frac{I_{En}}{I_E} = \frac{I_C}{I_C + I_B} = \frac{1}{1 + \frac{I_B}{I_C}} = \frac{1}{1 + \frac{G_B}{G_E}} = \frac{1}{1 + \frac{1}{\frac{W_E N_{DE} D_{nB} n_{iB}^2}{W_B N_{AB} D_{pE} n_{iE}^2}}} = \frac{1}{1 + \frac{W_B N_{AB} D_{pE} n_{iE}^2}{W_E N_{DE} D_{nB} n_{iB}^2}} \qquad 5.5.3.$$

For equal Homojunction:

$$\frac{n_{iE}^2}{n_{iB}^2} = 1 \text{ but in Heterojunction } \frac{n_{iE}^2}{n_{iB}^2} = Exp\left[-\frac{(E_{gE} - E_{gB})}{kT}\right] \qquad 5.5.4.$$

If Emitter Band Gap > > Base Band Gap then:

$$\frac{W_B N_{AB} D_{pE} n_{iE}^2}{W_E N_{DE} D_{nB} n_{iB}^2} \ll 1 \quad and\ \gamma \sim 100\% \qquad 5.5.5.$$

This proves that in Heterojunction diode there is a copious injection of majority carriers from Large Band Gap Material to Narrow Band Gap Material. This will help achieve much more efficient and more powerful LEDs as well as LASER Diodes.

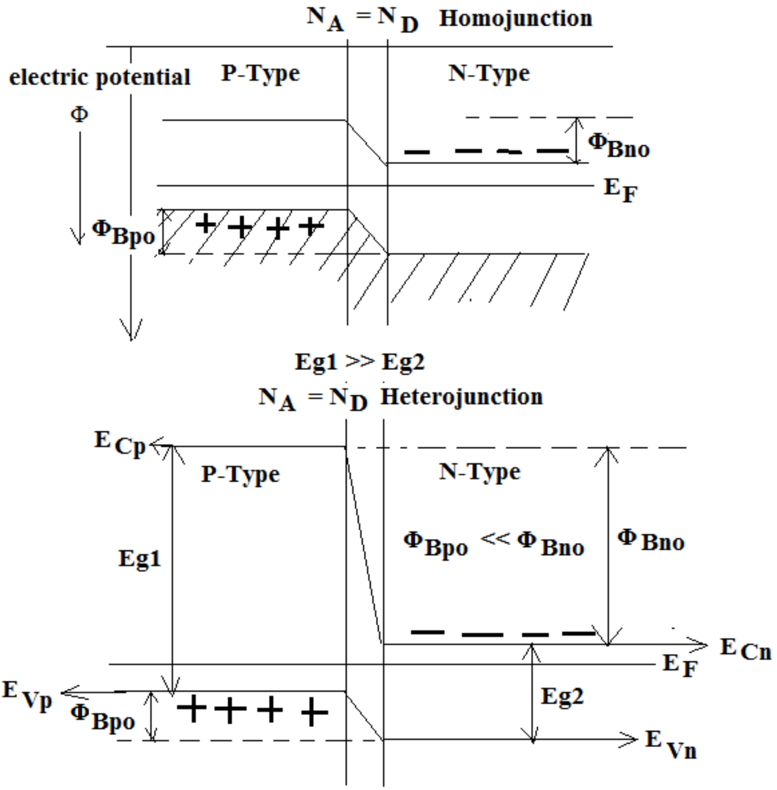


Figure 5.5.4. Comparative Study of Energy Band Diagram of Homojunction and Heterojunction diodes at zero bias.

Section 5.5.3. Double Heterojunction LED

Homojunction LED suffer from poor internal quantum efficiency and considerable loss of radiated photons. These problems are considerably overcome by Double Heterojunction LED.

As seen in Figure 5.5.5. P-Type lightly doped GaAs(Band-Gap = 1.42eV) is sandwitched between two large band-gap layers of AlGaAs(Band-Gap = 2.546eV).

In AlGaAs, Al and Ga are in Group III and As is in Group V. By controlling the stoichiometric coefficient ‘x’ in $Al_xGa_{(1-x)}As$, the band-gap can be tailored according to the requirement.

$$\text{The Band Gap of Ternary Compound} = E_g = (1.424 + 1.247x)eV \qquad 5.5.6.$$

At x = 0.9, the band-gap of $Al_xGa_{(1-x)}As$ is 2.546eV.

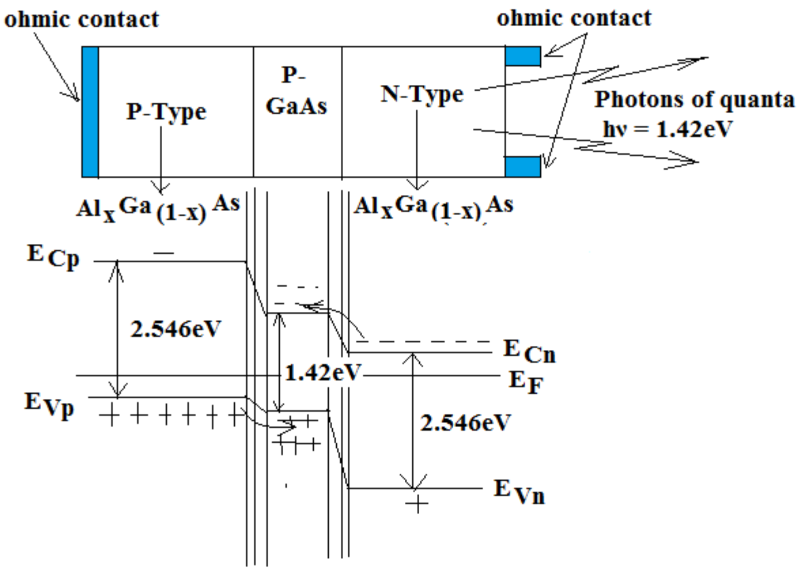


Figure 5.5.5. Device Structure of Double Heterojunction LED and its Energy Band Diagram.

As we see in Figure 5.5.5., a narrow band gap layer(GaAs, 1.42eV) is confined or sandwiched between two wide band gap layers(AlGaAs,2.546eV). In effect a thin quantum well is created in the middle layer. This causes heavy injection of holes from the P-Type layer on the left and heavy injection of electrons from N-

Type on the right. So a large amount of electrons and holes are confined in this narrow quantum well leading to rapid direct radiative recombination.

His has three distinct advantages over Homojunction Surface Emitting LED.

1. The interface state density at the lattice matched junctions of GaAs and AlGaAs is orderts of magnitude smaller than the free surface density of states at the emitting surface in homojunction case. This drastically improves internal Quantum Efficiency of the recombination taking place in the narrow Quantum Well.
2. Large Band Gap Bulk has zero absorption coefficient for photons of 1.42eV.Hence AlGaAs is completely transparent to outgoing photons thereby improving the external Quantum Efficiency.
3. The unequal Band-Gap of the two outer layers ensures almost 100% injection of majority carriers from both sides into the narrow Quantum Well.
4. Narrow Band Gap material has larger refractive index as compared to that of the adjacent layers. This helps confine the photons within the Quantum Well by total internal reflection. This light confinement helps increase the efficiency of stimulated recombination and emission process which is directly dependent

on the strength of the optical field in the Quantum Well. Initially there is spontaneous emission. This triggers stimulated emission. Hence in double-heterojunction LED, both spontaneous as well as stimulated emission takes place.

These four factors greatly improve the external as well as internal Quantum Efficiency. The Quantum Well comprises of 0.1 to 0.2μm thick narrow band-gap material. These surface emitting structures and edge emitting structures(not described here) are particularly useful in improving the coupling efficiency with te optical fibers.

Syllabus of EC1X61_IC Design and Fabrication Technology-Microelectronics [3-1-0]
This module gives the syllabus of EC_1X61. The lecture notes covering this syllabus is given as Chapter 6 of Solid State Physics and Devices_The harbinger of third Wave of Civilization.

EC1661_IC Design and Fabrication Technology-Microelectronics [3-1-0]

Introduction – 50 years journey of IC Technology from micro to nano era_a broad overview.

Bipolar Junction Transistor Fabrication Process Flow; substrate preparation, **oxidation process** (wet and dry), **photo-lithography**, **diffusion** (pre-deposition and drive-in), buried layer diffusion_drive-in, **Chemical Vapour Phase Deposition of thin Si epitaxial layer**, isolation diffusion, base diffusion, emitter and collector pre-deposition diffusion, **metallization** for contact pads and interconnections, scribing of individual dies, mounting on ceramic wafers, ceramic wafers are bonded to headers, hermetic shielding and packaging (TO5 or Dual-in-Line package).

Diode, Resistance and capacitor realization by IC Technology.

BJT IC layout consideration.

MOS instability and C-V measurement curves under steady state DC voltages, ideal High-Frequency Capacitance and Low Frequency Capacitance.

Theoretical formulation of Threshold Voltage in MOS.

Comparative studies of BJT and MOS Technology.

*Complete CMOS Process Flow:***Ion –Implantation**, **LOCOS**, **Shallow Trench Isolation**, Field –implant under LOCOS, Buried and Epitaxial layer option, LPCVD of Poly-Si gate, self-aligned Source and Drain, side-wall spacers along the edge of gate, source/drain LDD, **multi-level interconnection**-back end processing.

Using ATLAS and ATHENA carrying out the following simulations;

PN Junction Diode Simulation.

NPN BJT Simulation.

CMOS simulation.

Text Book: . Uploaded on http://Cnx.org SSPD_Chapter 6 by Bijay Kumar Sharma;

References: ”Silicon VLSI Technology: Fundamentals, Practice and Modeling”, Plummer, Deal and Griffin, Prentice Hall.

Chapter 6. IC Fabrication Technology from 60s to date_Part 1

Part_1 of Chapter_6 of SSPD deals with the fabrication steps of Motorola First Generation Operational Amplifier IC Chip MC-1530. This was the first analog IC Chip introduced in 1965. Until then only RTL and DTL chips were in market.

CONTENTS

- Introduction.
- Two Broad Categories of I.C.
- Planar Fabrication Processes.
    - Epitaxial Diffusion Process.
    - Mounting on header and packaging.
    - Photolithography.
- Fabrication of MC 1530.
    - Preparation of single crystal silicon wafer.
    - Thermal Oxidation.
    - Technology of LOCOS.
    - Wet etching vis-à-vis dry etching.
    - Poly-Si buffered LOCOS.
    - Buried Layer Diffusion.
    - Passivation.
    - Solid State Diffusion.
        - Constant Source Diffusion.
        - Limited Source Diffusion.

        - Diffusion Systems.
        - Buried Layer Diffusion.

    - ○ N epitaxial layer growth on P substrate.

        - Epitaxial Layer Growth by Chemical Vapour Deposition(CVD).

    - ○ P type Isolation Diffusion for electrical isolation of IC components.
    - ○ P type Base diffusion.
    - ○ N+ diffusion for Emitter and Collector Contact.
    - ○ **Aluminum Contact Pads and Aluminum Interconnections of IC components.**
    - ○ **Flow-Chart of the processing of MC-1530**

Chpater 6. Integrated Circuit Design and Fabrication Technology

Section (6.1) Introduction:

In 1930 the band theory of solids was presented by F. Block, A. H. Wilson and others. This was the beginning of the science of semiconductor technology but for next 20 years no solid state device found any wide application in the field of Electronics Engineering. It was the invention of transistors by Bardeen, Brattain and Shockley (1948) that raised the stature and importance of Solid State Technology in the field of device fabrication. The American Multinational Electronic companies with their huge in-house Research and Development facilities made important breakthrough and remarkable innovations in Solid State Technology resulting in newer processes and techniques such as diffusion , ion-implantation, Deep Ultra-Violet Lithography, poly-Silicon emitter contact and low temperature epitaxial growth. Development of silicon and silicon based planar technology permitted the fabrication of transistors, diodes, capacitors, resistors, and their interconnections on a monolithic single crystal chip. Integrated Circuit (I. C.), technology helped microminiaturize complex electronic circuit, considerably cut down the power supply requirement, by dispensing away with wire interconnections reliability was considerably increased and by mass producing standardized chips the cost per chip was considerably reduced. Apart from these advantages, I. C. Technology gave very good temperature tracking of devices (active or passive) lying on same isothermal of the I.C. chip (temperature tracking results in favorable performance feature), gave matched transistors which are indispensable in I.C. design philosophy and gave higher speed by reducing the parasitic capacitances and by reducing the propagation delay. But I. C. technology limits the range of component values (resistors and capacitors) which can be monolithically realized and has no provision for close tolerance components and for inductors. The limited repertoire of component types and values put some constraint on the circuit designer. This constraint has been partly removed by the use of hybrid circuits and partly by a wider application of CMOS devices.. In hybrid circuits the I. C. chips are used in conjunction with thin film and thick film components which gives much greater flexibility to circuit designers especially in designing Analog Circuits (or linear/nonlinear circuits). In Analog circuits proliferation of I. C. technology has been much less because of lack of standardization in analog applications and because of very specialized requirements. Thin Film and Thick Film Technology afford much closer tolerances and much wider spectrum of component values.

In first phase of I. C. Technology where Photolithography was used, a resolution of 10 μm could only be achieved. This put a very low limit on the system complexity which could be integrated within the chip mainly because of a low packing density. Hence only circuits of a fair degree of complexity could be integrated. This was known as Small Scale Integration (S. S. I.).

But with the development of Deep Ultra Violet Lithography and Ion Beam Processing (such as Ion Beam Milling and Ion Implantation), which afford submicron resolution and with the development of MOS technology, much higher packing density has become physically realizable. This has greatly raised the limit on system complexity which can be Integrated. Now it is possible to integrate a fully functional block. This level of system complexity when integrated gives rise to Medium Scale Integration (MSI) chips, Large Scale Integration (LSI) chips and Very Large Scale Integration (VLSI) chips.

Examples of S.S.I., M.S.I., L.S.I. and V.L.S.I. are given in Table (6.1).

Table (6.1) SSI, MSI, LSI and VLSI/ULSI Chips.

	S S I Quad 2 input SN 5400	M S I Decade Counter SN 5490	L S I 1024-bit Shift register and 4K(4096 bit) memory elements	VLSI/ ULSI Intel8086/ Pentium IV
Chip area(sq. mils)1 mil = 25 microns	3000	5750	-	½ cm x½ cm
Number	36	102	More than	10^5 to40

of components			100 gates	million CMOS

The ultimate packing density achieved to this date (2011) is 1.3 billion devices in dual core Itanium 2 processors chip manufactured by INTEL. These are examples of Ultra Large Scale Integration (ULSI). LSI techniques have made possible small modern hand held calculations, microprocessor based microcomputers, powerful digital computers in relatively small sizes and the integration of much of the electronics within audio amplifiers and televisions sets.

As of 2005, a IC Chip fabrication foundry facility costs over $1 billion . It has the following features:

- Wafers are 300 mm diameter;
- 65nm technology is gradually being replaced by 45nm technology;
- Deep Ultra Violet (193nm) immersion Lithography is being used;
- Aluminum interconnects are being replaced by copper interconnects to overcome the problems associated with lateral scaling;
- Low-K dielectric insulator is being used for the Cu-interconnects overlay surface so as to maintain a high velocity of prorogation;
- Silicon on Insulator is being used to reduce the parasitic capacitances;
- Strained Silicon directly on Insulator is being

used for higher electron mobility in 2D channel.

The development of Liquid phase and Vapor phase epitaxy has made Gallium Arsenide technology commercially feasible. Metal Organic Chemical Vapor Deposition (MOCVD), a low cost technology, and Molecular Beam Epitaxy, Ultra high vacuum technology, are further developments in the direction of GaAs technology.

In Table (6.2) a comparative study of GaAs , Si and Graphene is made.

Table (6.2) A comparative study of GaAs, Si and Graphene.

	Band Gap300K	Mobility of electrons300K_cm2/(V-sec)	Mobility of holes300K_cm2/(V-sec)	Intrinsic ConcentrationAt 300K(number/cc)
Si	1.12eV(Indirect)	1500	600	1.45×10^{10}
GaAs	1.43eV(direct)	8500	400	9×10^{6}
Graphene	Zero-gap semiconductorOr semi-metal	200,000	No holes	$10^{12}/cm^2$

Comparative study of Silicon,Gallium Arsenide and Graphene

Table (6.2) A comparative study of GaAs, Si and Graphene.

As seen from the table, mobility of electrons is five times higher in GaAs than in Si. Also ni (intrinsic concentration) is much lower in GaAs as compared to Si. Hence GaAs based I.C. chips are going to be five times faster than Si based I.C. chips and by proper dimension sealing very high packing density can also be achieved. Thus GaAs I.C.'s seen to hold very good promise for Gigabit/Gigahertz applications. Defense department of U.S.A. is going in a big way for GaAs based Very High Speed Integrated Circuits (VHSIC) which will realize real time signal processors for weapon systems of the next decade.

As seen from Table (6.2), Silicon has indirect band gap while GaAs has direct band gap. As a result, electron-hole pair recombination is accompanied with phonon emission in Si. That is direct transitions are non-radiative. Hence GaAs based systems are indispensible for Opto-Electric applications. Moreover by proper compensation of GaAs very high resistivity substrate 10ˆ8 ohm-cm can be achieved which is directly applicable in Microwave Integrated Circuits (MIC).

From Table (6.2), we also see that graphene will emerge as the material of the future. It is 2D carbon layer where carbon are bonded in honeycomb chicken net structure. It is a semi-metal or zero-gap semiconductor . Theoretically the mobility is very high at 200,000cm2/V-sec but in practice the substrate phonon scattering limits it to 20,000cm2/(V-sec).

In the era of discrete component based Electronic Systems there were three distinct and non-interacting disciplines: device design, circuit design and system design. But with the advent of I.C. based Electronic Systems the traditional boundaries between the three disciplines have been obliterated. Today an electronic circuit designer or I.C. chip designer must be knowledgeable in semiconductor fabrication technology, in the physical electronics and circuit modeling of semiconductor devices and in the methods and techniques of circuit analysis and design. Thus a chip design must be aware of the constraints and opportunities of device technology and circuit theory as well as must have a sufficient knowledge of Systems Engineering so that he is aware of the requirements of the users of his products. By the same token the systems Engineer must be aware of the capabilities and limitations of the present day State of Art of I.C. technology and accordingly limit his requirements expected from I.C. functional blocks currently available in the market.

Today the system’s design is fully Computer-Aided Design. In Digital System Design, most complex systems can be described by Hardware Description Language and implemented on Field Programmable Gate Array. Similarly there are softwares for designing Analog and Mixed Signal Systems.

Section (6.2) Two Broad Categories of Integrated Circuits

Integrated circuits broadly fall in two groups:

Monolithic Circuits and Hybrid Circuits.

A monolithic circuit is fully realized on a single silicon chip. Through Planar Fabrication Processes all circuit elements are simultaneously fabricated and interconnected on the same silicon chip. This monolithic implementation is very convenient for batch production hence is very suitable for low cost mass production. The range of passive component values, their tolerances and their temperature coefficients is more limited than with thin film processes but quite adequate for most digital circuits and some analog circuits. By CMOS technology, digital system as well as analog system implementation has been made feasible. CMOS enables the realization of resistances and capacitances by proper configuration of NMOS itself. If a greater variety of resistances are required then as is required in case of Active Filters we resort to switch-capacitor filter technique. This enables us to achieve a wide range of time-constants.In fact a large number of off-the-shelf-digital integrated circuits as well as numerous "custom" (special purpose) circuits for large-quantity production contracts are manufactured in this manner.

A hybrid circuit is essentially the marriage of monolithic technology and thin film technology realizing the best of both the technologies. Thin film technology and thick film technology offer greater design freedom with respect to physical realization of passive components, resistors and capacitors, in the sense that greater range of component values, closer tolerances and lower temperature coefficients can be achieved.

There are two variations of this hybrid technology. In one variation thin film or thick film passive components are realized on a suitable substrate (such as coming glass or ceramic) and these are interconnected among themselves and to the discrete active components (diodes and transistors) situated on the same substrate by metalized interconnecting patterns. Thus the complete working circuit is achieved on the same substrate. This is strictly known as hybrid circuit.

In another variation, the active components are formed within a silicon chip and the passive components pattern is deposited by thin film techniques on top of the silicon dioxide passivating layer covering the active circuit. These are known as compatible circuits which is an extension of monolithic art.

In this Chapter we will concentrate on the state of art of Monolithic Technology.

Section (6.3) Planar Fabrication Processes for Integrated Circuits

In I.C. technology all the fabrication steps are performed at the surface of a silicon crystal and all contacts lie in one plane hence these processes are called Planar Fabrication Processes. Planar fabrication process consists of three routes for achieving monolithic circuits:

(a) The epitaxial diffusion process,

(b) The diffused-collector process, and

(c) The triple diffusion process.

Here we will describe epitaxial diffusion process which has become the standard method for fabricating IC Chips.

6.3.1 Epitaxial Diffusion Process:

Epitaxial diffusion process consists of the following steps,

(a) preparation of single crystal silicon wafer,

(b) growth of epitaxial film over the silicon wafer,

(c) growth of silicon dioxide film,

(d) photo masking and photo etching ,

(e) impurity diffusion/ion implantation,

(f) vacuum evaporation of Aluminum.

Some of the above steps are employed more than once in the sequence.

6.3.2 Mounting on header and packaging.

After these steps the wafer is scribed with a diamond-tipped scribing tool and separated into individual chips or dies. These individual silicon chips or dies are mounted on ceramic wafers by means of solder glass .The ceramic wafers are attached to suitable header by means of a high temperature eutectic solder . Or the die can be mounted in all ceramic flat package . Gold Wires only 0.001 inch in diameter are bonded from the circuit to the proper package leads by thermo-compression or ultrasonic bonding.

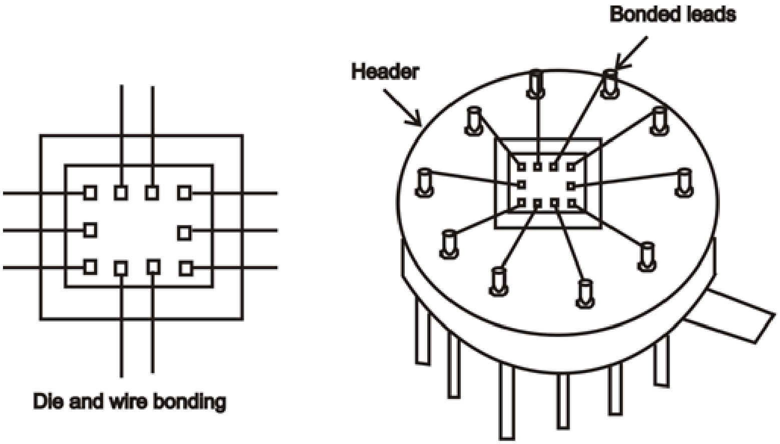


Figure 6.1 The individual die or chip mounted on the header with leads bonded in place.

The header is hermetically sealed by a suitable packaging material which is impervious to moisture and gas. The packaging should satisfy the following requirements :

1. It must be mechanically strong to withstand the stresses during use;
2. It should be small and of a shape which permits compact aggregation ;
3. It should provide easy electrical connections from inside circuit to the outside world;
4. It should keep parasitic inductance and capacitance to a minimum;
5. Electrical insulation of the chip and the leads with respect to header should be maintained and yet the thermal resistance from the chip to the ambient should be as low as possible;
6. The package must maintain an interior

environment which is stable and congenial to the circuit;

7. And lastly the package is called upon to shield the circuit from light and occasionally from magnetic fields. In space applications even nuclear radiation shields are now feasible.

Two basics packages currently in wide use for I.C.'s are : new flat package and other is a multileaded TO-5 package . Flat packs can be mounted on both sides of a printed-circuit board whereas TO-5 packages can be mounted only on one side of PCB because of their plug in nature. Flat package gives advantages in compactness of 5 or 6 to 1. The two are illustrated in Fig.(6.2)

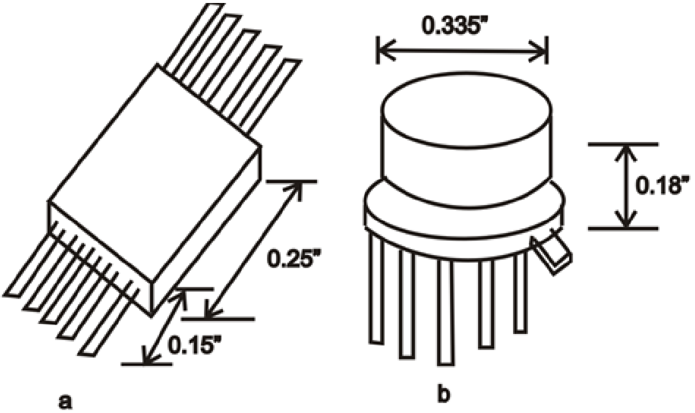


Figure 6.2 A comparison of the (a) flat package and the (b) 10 lead TO-5 package

The six steps of Epitaxial Diffusion process will be briefly described and illustrated by actually going through the fabrication steps of MC 1530, the first generation of high performance operational amplifier marketed by MOTOROLA in mid-60’s.

The heart of Epitaxial Diffusion is the basic masking technique (or photolithography) which enables the selective impurity diffusion or ion implantation into silicon wafer and which enables to achieve a definite pattern of Aluminum interconnections according to the circuit requirement.

6.3.3 Photolithography:

To fabricate a complete integrated circuit a set of photo masks are used in sequence. Each photo mask has a pattern according to the specific requirement of the step in which it is going to be used. Furthermore the image produced by each mask must properly align with the image of the preceding mask so that the sum total effect of all the masks results into the desired circuit.

Photo mask fabrication consists of the following steps:

1. Preparation of Master Artwork:

The first artwork is prepared 200(×) the actual integrated circuit layout. This makes the preparation of the artwork convenient as well as accurate. The master artwork is cut on Rubylith by coordinatograph. Rubylith is Mylar laminate with a pealable opaque overlay. By cutting the overlay with a knife edge it may be removed to form clear areas in an opaque background.

1. First reduction:

Once the initial artwork is prepared, the single image is reduced 20:1. This first reduction image must be an extremely accurate reproduction of the master artwork therefore the photography should be sharp.

1. Printing of multiple array of images:

By a step and repeat camera, the first reduction image is contact printed at a point and then moved from that point to another point at an accurate spacing. This multiple printing continues until the desired array is achieved. By multiple-lens camera also this multiple array of images can be achieved.

1. Final reduction:

The multiple images formed are then reduced to the final desired size by producing a microphotograph of the print from the step-and-repeat machine. The final reduction is 10:1 and sharpness of focusing has to be maintained.

There is nothing sacrosanct about the 200(×) magnification of the master artwork and 20:1 first reduction and 10:1 second reduction. There are many considerations which decide the magnification and the reduction ratio. Also the sequence of steps described above is not strictly followed in preparation of photo masks. The formation of multiple images may precede, coincide with or follow the final image reduction.

The set of photo masks required for the fabrication of MC 1530 are:

1. Mask for n+ diffusion for buried layer;
2. Mask for p diffusion for isolation islands;
3. Mask for p diffusion for base and resistors;
4. Mask for n+ diffusion for emitter and collector contact;
5. Mask for aluminum contact windows;
6. Mask for interconnection aluminization;

These masks have been described in the sequence in which they will be used during actual fabrication. These masks are used to impart selectivity to different processes of fabrication by photo-resist process or photolithography.

The photoresist process consists of the following basic operation:

(1) A uniform layer of light sensitive photo resistive emulsion is coated over the wafer by a spinner machine.

(2) In a mask alignment and ultraviolet exposure machine, the photo-mask used, corresponding to the

fabrication step in question, is correctly aligned with respect to the previous pattern on wafer. Next photo-mask is brought in intimate contact with photo resist layer and the photo resist layer is exposed to intense ultraviolet light for a given stipulated time through the photo-mask. The KPR layer is baked in an oven at 100°C to dry it. As the lateral feature size has decreased from generation to generation, so has the exposure method changed and the wavelength of the exposing light has been progressively decreased so as to satisfy the Rayleigh Equation. In 1997, 0.35μm Technology used 248 nm Deep Ultra Violet light. As we moved to 0.15μm Technology we used 193μm DUV light. Today as we adopt 0.045μm Technology we are using DUV 193 μm immersion Technique. The Contact Printing gave way to Proximity Printing and Proximity Printing gave way to Projection Printing. Today we are using Deep UltraViolet (193nm) Immersion Technique. [For detailed discussion see "Journey of IC Technolgy from micro to nano Era- Part 3".]

(3) By photo resist developer, the image of the mask is developed in the photo resist layer . In positive photo resist the exposed layer gets removed leaving the wafer surface bare in those areas. In negative photo resist the unexposed layer remains soft and is dissolved away by the developer.

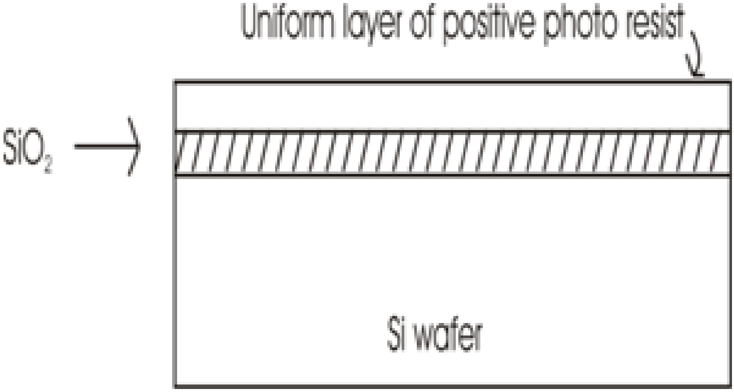
(4) By baking the wafers in baking oven at a given temperature for a given time, the remaining photo

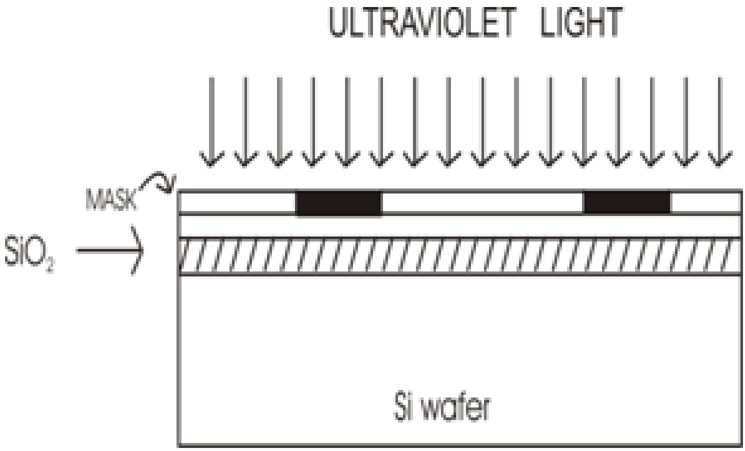
resist pattern on the wafer is fixed and becomes resistant to corrosive etches.

(5) Depending on the fabrication step, either SiO_2 or Aluminum layer on the substrate (or wafer) is etched through the photo resist pattern carving out an identical pattern on the surface of the substrate. Hydrogen Flouride (HF) is used for etching SiO_2 and Sodium Hydroxide (NaOH) is used for etching Aluminum layer. Hydrogen Fluoride is highly corrosive and is stored in Teflon beaker. Teflon tweezers are used to handle the wafers and Teflon gloves have to be used by the worker to protect himself from Hydrogen Flouride.

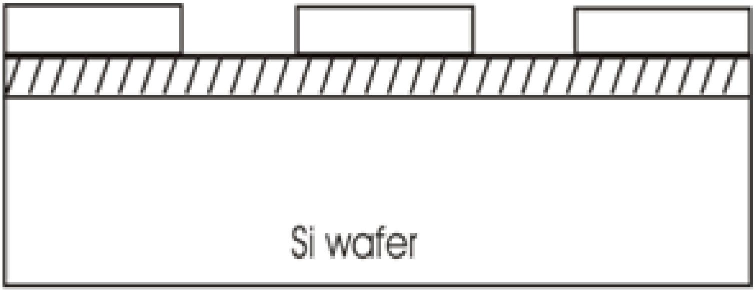
(6) Next by stripper ,the hardened photo resist pattern is stripped and the wafer arecleaned for the next fabrication step.

Photolithographic processes are illustrated in Fig (6.3)

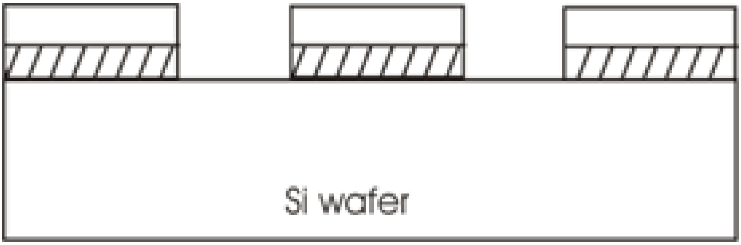




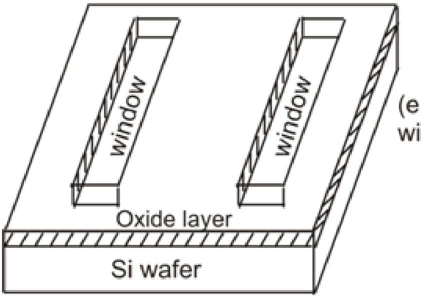
(a)Deposition of uniform layer of photo resist.
(b)Mask alignment and ultraviolet exposure.



(c) Photo resist is developed and fixed.



(d) pattern identical to that of the photo mask is etched out in SiO_2 and subsequently photo-resist is stripped.



(e)after stripping ,the resultant window pattern etched out in SiO_2

Figure 6.3. Different steps of Photolithographic Process.

Section (6.4). The Fabrication Steps of MC1530.

(all fabrication steps will be illustrated by figures at the end of this section)

6.4.1 Preparation of single crystal silicon wafer (This is already described in Chapter 2 of SSPD):

A single crystal silicon of given type , of given conductivity , of high crystalline perfection and having very clean surfaces is the starting point of I.C. fabrication.

The basic starting material is silicon in the highly pure form but polycrystalline. This high grade silicon is melted in a crucible. To this silicon melt a suitable n or p type impurity is added to impart the required type of conductivity. The maximum conductivity which can be imparted is limited only by the solubility of the desired impurity in silicon .A seed of single crystal (small, highly perfect and of desired crystallographic orientation) is immersed into the melt and gradually withdrawn .By this crystal pulling , a nearly perfect single crystal of about 1 inch. in diameter and 8 inches to 10 inch in length is pulled .The finished crystal ingot is frequently ground to an uniform diameter by center less grinding . By X-ray techniques, the crystal is correctly oriented according to crystallographic planes and then sawed into wafers using a diamond-impregnated blade .The wafers , after being sliced , are mechanically lapped on abrasive lapping machines (the abrasive being Al_2O_3 or diamond) to the thickness required for I.C. which is 6 mil i.e 150 microns thick. The lapped wafers are then polished to a mirror finish with graded diamond polish, the final polish being done with quarter micron diamond paste .The lapping and polishing operations result into many surface imperfections on the wafers .These imperfections are removed by chemical and gas etching technique.

Table 6.3 Identity marks of the Wafer to identify its orientation and semiconductor type.

α(angle between primary and secondary flats)	Type	Orientation
45°	N	<111>
90°	P	<100>
180°	N	<100>
0°	P	<111>

For MOS fabrication, wafers with crystal orientation <100> are used. This helps achieve a lower threshold voltage. For BJT <111> orientation is preferred.

A single crystal wafer of high crystal perfection and surface cleanliness is obtained. For MC 1530 (or for that matter for any I.C.) we start with a p type silicon wafer which is 6 mil thick and has a resistivity (ρ) of 10 ohm-cm. This corresponds to a doping density of Na =1.4 x 10^{15} (cm^{-3}). The silicon wafer will be 200mm in diameter and may contain several hundred potential dies or chips each containing a complete integrated circuit. In 2010, the wafer size is being increased to 300mm as the circuit complexity increases. In Figure (6.4) a typical wafer with chequer-board pattern drawn on it is shown .The black square on the chequer board pattern represent the potential die or chip. In one run, hundreds of Si wafer are simultaneously processed and each of the wafers contains hundreds of I.C.Chips. This amenability of IC technology to batch production thereby leading to mass production of I.C. chips greatly reduces the cost per chip which is a big step in making the applications of I.C. chips universal and common place and allowing the I.C. based system to penetrate every sphere of life , such as optimizing the industrial processes on one hand and on the other liberating the housewives from mundane daily chore of household works for more useful and satisfying social works .Thus I.C. based systems are making electronics play progressively a more important , perhaps crucial , role in making a secure future of mankind .

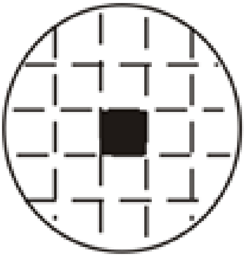


Figure 6.4 A silicon wafer where a potential die has been indicated by a black square.

The resistivity of the p type substrate is kept very high (10 ohm-cm) because all the components on a single chip will be isolated from one another by being kept reverse biased with respect to the substrate . Since substrate is p type it will be kept at most negative potential “-VEE “. And the most positive potential of the circuit is “VCC”. Therefore maximum reverse potential with respect to substrate will be (VCC + VEE). Hence substrate breakdown voltage must be high and at the same time the parasitic coupling of each component to the substrate must be kept at a minimum so as to derive optimum high frequency performance out of the chip. Both these objectives are achieved by having high resistivity.

6.4.2 . Thermal oxidation :

Next important fabrication step is thermal oxidation which enables the growth of a uniform thin skin of

SiO_2 , approximately 3,000A_o to 10,000A_o thick (0.3 to 1 μm), over the clean silicon surface.

$3000 \times 10_{-10}$m $= 0.3 \times 10_{-6}$ m $=$ 0.3um(micro meter) $=$ 300nm.

Silicon dioxide plays four important roles in I.C. fabrication namely acts as i. mask for diffusion, as ii.Gate Oxide in MOSFET, as iii. Field Oxide for electrically isolating the MOSFETs and for iv. passivation.

At 45nm technology, the mask is typically 100nm thick, Gate Oxide is 40nm thick, field Oxide is 200nm thick and passivation layer is 300nm thick.

SiO_2 layer for Masked diffusion or ion implantation – Most of the dopants or impurities commonly used for doping have very low diffusion coefficients in SiO_2 as a result even a long exposure to dopant (or impurity) atoms at high temperature no impurities penetrate SiO_2 . This property enables one to etch apertures of selected shape and size in SiO_2 layer by the use of photolithographic techniques and then the wafer is exposed to the diffusion at given temperatures for given time. Because of masking property of SiO_2 only selective diffusion through the aperture into Si wafer takes place. Thus by masked diffusion , using SiO_2 mask suitably, the conductivity properties of the wafer are modified in desired localized region in a controlled fashion.

P type dopants (group III elements) are boron(B), Aluminium (Al),gallium(Ga). Boron has very slow diffusion in SiO_2 but rapid diffusion is Si whereas Gallium has rapid diffusion in both. On the other hand Aluminum has very low solid solubility in Si (maximum surface concentration is only 0.002 percent) therefore Al used only for ohmic contact pads. Hence in masked diffusion, Boron is universally used as P type diffusant.

N type dopants are Phosphorous (P) and Arsenic (As). Both are effectively masked by SiO_2 but 'As' has very low diffusion constant in Si. Therefore Phosphorous is universally used as N type diffusant . B and P have nearly equal diffusion constants in Si and they both have very high solid solubility, permitting a high surface concentration.

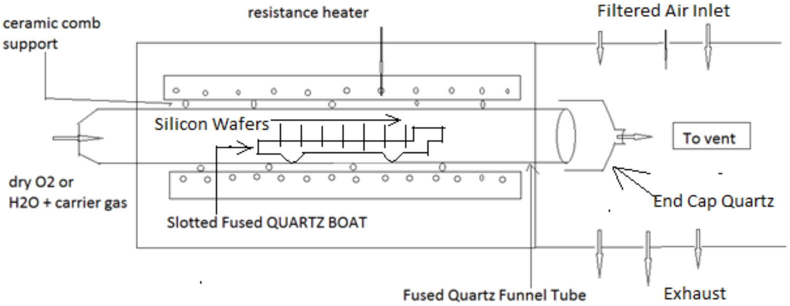


Figure 6.5 Schematic cross-sectional view of a resistance heated oxidation funace.

[The silicon wafer loading area is always in the laminar hood. The laminar flow hood reduces dust and particulate matter in the air surrounding the wafers and minimizes contamination during wafer loading]

KINETICS OF GROWTH:

DRY OXIDATION

Si (solid) $+ O_2 \rightarrow SiO_2$(solid)

WET OXIDATION

Si(solid) $+ 2H_2O$ (steam) $\rightarrow SiO_2$ (solid) $+ 2H_2 \uparrow$.

Dry Oxidation is much slower than Wet Oxidation as is evident from the Table 6.4 and Table 6.5.

Table (6.4) Oxide Growth (μm) versus temperature and time for dry oxidation for

< 1 0 0 > crystal orientation.

Temp Time (hr)	800°C	900°C	1000°C	1200°C
0.3 hrs			0.023	0.094
1.0 hr		0.018	0.05	0.195
3.0 hrs	0.013	0.038	0.10	0.350
10.0 hrs	0.03	0.087	0.23	0.680

30.0 hrs	0.065	0.10	0.40	1.100

Table (6.5) Oxide Growth (µm) versus temperature and time for wet oxidation for

< 1 0 0 > crystal orientation.

Time (hr)	900$_o$C	1050$_o$C	1250$_o$C
0.3 hrs		0.27	0.54
1.0 hr	0.23	0.55	1.00
3.0 hrs	0.45	0.95	1.80
10.0 hrs	0.93	1.80	3.50

This oxide layer is called thermal oxide and it uses a part of the silicon surface atoms . Hence 44% of SiO2 layer is below the silicon surface and 56% above the original silicon surface as shown in Figure 6.9.

Temperatures of 1000$_o$C and 1 hour thermal oxidation time are the typical oxidation parameters. Where thick layer (0.5 µm or 500nm thick) of SiO_2 required one goes for WET OXIDATION and where thin layer (30 nm-100nm) accurately controlled is required there we go for DRY OXIDATION.

(1) Dry oxidation is slow but gives UNIFORM , RELATIVELY FREE FROM DEFECTS and ELECTRICALLY STABLE thin layer (30 nm to 100 nm) as required in MOSFET Fabrication . MOSFET requires insulating gate oxides(5 to 50 nm thick) and it requires 200 nm thick field oxide.

Initially because of instability and latch - on problems of MOSFET , high packing density which is afforded by MOS circuits was technically unfeasible. Eventually oxide was doped with phosphorous and P glass achieved . This anode P glass is impervious to Na_+ ions migrations and MOSFET was stabilized. This led to the development of VLSI and ULSI fabrication. (Andy Grove of Intel fame had an important role in developing the MOS Technology).

(2). Wet oxidation- This is carried out in steam ambient and as it is clear from the Table (6.4) it accelerates the growth rate of SiO_2 but it is not as free from microscopic defects as is dry thermal oxides. Hence wet Thermal Oxides is primarily used as MASKS Application during diffusion and wet oxides are also used as Field Oxides.

In both types of thermal oxide, interface (Si/SiO_2) trap densities are minimized.

The oxidation is carried out in oxidation furnace as shown in Figure (6.5) where loading of the wafers is done in Laminar Flow Hood. This minimizes contamination and dust particles on wafer surface.

The oxidation temperature is 900$_0$ to 1200$_0$C and the typical gas flow rate is 1 cm/sec. The oxidation system uses microprocessors to regulate the gas flow sequence, to control the automatic insertion and removal of silicon wafers, to ramp the temperature up (i.e. to increase the furnace temperature linearly from a low to the oxidation temperature so that wafers will not wrap due to sudden temperature change), to maintain the oxidation temperature to within ±1$_0$C and ramp the temperature down when oxidation is completed.

Initially the growth rate over a fresh silicon surface is reaction controlled since for thin layer the growth is decided by the reaction rate of reacting Si and O_2.

In Reaction Controlled Oxidation

x =

$$\frac{B}{A}$$

(t + T) - linear law (1)

B/A = linear rate constant.

Where A, B and T are constants

$X =$ thickness of oxide, $t =$ time of oxidation.

Definition of thin layer is: $X << A/2$

$$\frac{4B(t+T)}{A^2}$$

$<<1$ - (2)

When oxide layer is thick i.e. $X >> A/2$ (long duration oxidation)

$$\frac{4B(t+T)}{A2} >> 1$$

then $X =$

$$\sqrt{B(t+T)}$$

→ parabolic law of growth (3)

Here the rate is parabolic and $B =$ parabolic constant .

Because of initial linear law of growth and finally parabolic law of growth during oxidation over a silicon surface with an initial oxide step , the step gets invisible and after long oxidation duration we get a smooth plane surface as shown in the Figure

(6.6).

This is very desirable feature for I.C fabrication because metallic inter connection are over the oxide surface covering the chip . The removal of undulations by long duration oxidation makes metallic interconnection reliable and of uniform strengths .

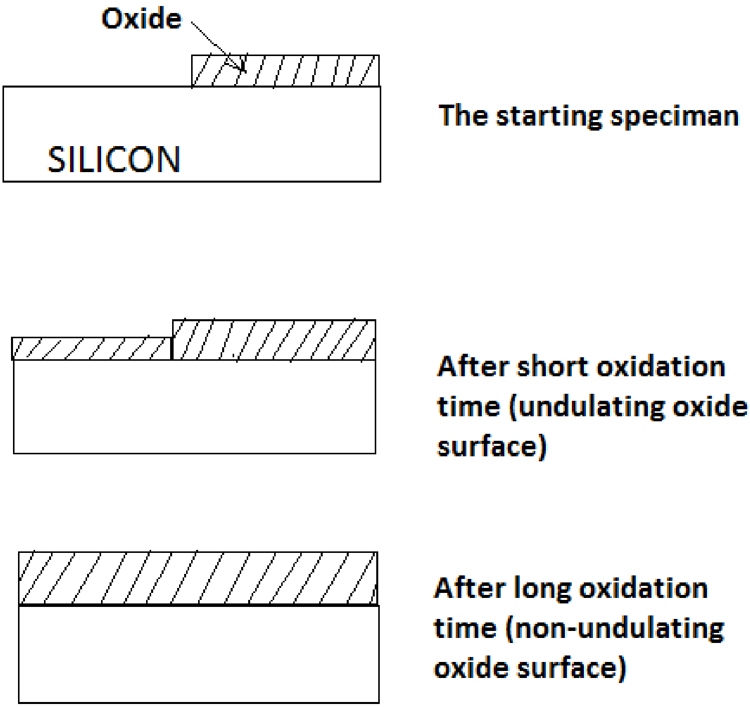

Oxide
SILICON
The starting speciman
After short oxidation time (undulating oxide surface)
After long oxidation time (non-undulating oxide surface)

Figure 6.6. Over the total fabrication cycle the linear and parabolic law compensate each other to provide a planar oxide surface which is ideal for depositing Al layer and etching out the interconnections.

By putting the wafer in oxidation furnace we get an uniformly thick oxide layer.

Typical furnace cycle is 15 minutes at 900°C in an H_2O ambient. This produces 40nm thick oxide layer. The usual practice is to make H_2 and O_2 react in the back end of the furnace to produce steam for H_2O ambient. This is a cleaner method for producing steam.

This same 40nm could be achieved by a furnace cycle of 45 minutes at 1000°C in dry oxidation ambient

Bruce Deal and Andy Grove developed linear-parabolic model for thermal oxidation. This is referred to as Deal-Grove model. It is used in SUPREM software developed by Stanford for process simulation.

Often we need a local thick oxide layer as is the case in CMOS technology. This thick local oxide layer is called field oxide and it helps electrically isolate two adjacent MOS devices. This thick local oxide layer is achieved by a technique called LOCOS (LOCal Oxidation of Silicon).

6.4.3 .Technology of LOCOS

As shown in Figure (6.7), 40 nm thick Oxide Layer is grown in Oxidation furnace and in a second furnace 80 nm thick Silicon Nitride layer is grown by Low Pressure Chemical Vapour Deposition (LPCVD) method.

$4NH_3 + SiH_4 = Si_3N_4 \downarrow + 12H_2 \uparrow$

This reaction is carried out at 800ºC below atmospheric pressure. This helps a better uniformity of deposited films over large lots of wafers.

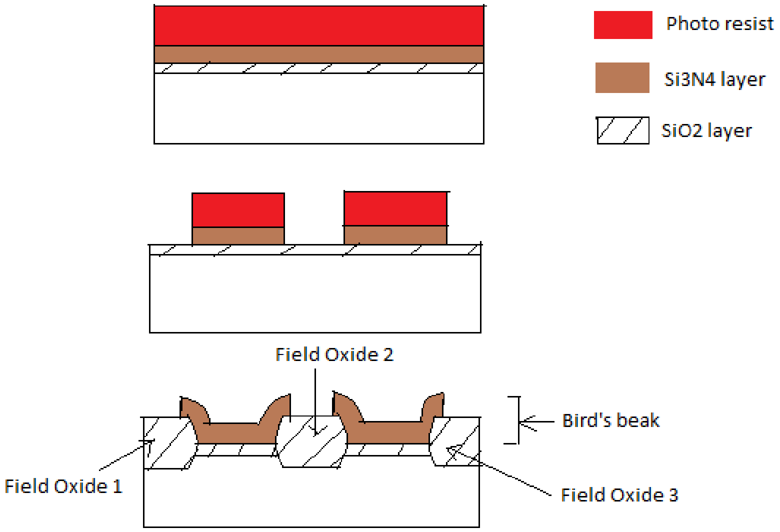
Si3N4 is covered by KPR and exposed through a suitable mask. We use positive KPR so that exposed portion remains soft and it is dissolved by the developer. The portion where local oxidation is to take place there KPR is dissolved and the portion where no local oxidation is required there the KPR is retained and fixed. Through the aperture in KPR film, dry etching of Si3N4 is carried out in Fluorine Plasma Chamber. As the feature size has decreased due to lateral and vertical scaling we have moved from WET PROCESSING to DRY PROCESSING.

6.4.4 Wet etching vis-à-vis dry etching

Wet etching is chemical etching where etching takes place in vertical and lateral direction also as shown in Figure 6.8. This is known as undercutting and causes the resultant pattern to be of larger dimensions as compared to that on the mask. To avoid this as we move to smaller feature size we prefer dry etching which is also known plasma etching, sputter etching, reactive ion etching (RIE).

Wet etching is isotropic hence undercutting results but is selective. Dry etching is anisotropic hence there are vertical edges and the resultant pattern is an identical replica of the pattern on the mask but it is non-selective. Hence dry etching effects the mask as well as the underlying layer of the target. In actual practice a compromise has to be made between the two. This compromise is achieved in RIE. Here we have neutral reactive specie and ionized reactive specie. Neutral specie chemically reacts with the target and Ionized specie reacts physically with the target. So we achieve anisotropic etching and selectivity also.

Si3N4 is etched out as shown in the middle diagram of Figure 6.7. In the remaining portion Si3N4 acts like a shield to oxygen atoms and does not allow local oxidation. Once the desired apertures have been formed in Si3N4 layer the fixed KPR is stripped. Next the wafer is put in Oxidation furnace to create Field Oxide 500nm thick through a furnace cycle of 1000°C for 90 minutes in Steam ambient. Field Oxide is formed where underlying oxide pad is exposed to steam. Where Si3N4 is present, no oxidation takes place as shown in the bottom diagram. But we also see the growth of oxide layer underneath the Si3N4 layer at the periphery of oxide layer. This gives rise to Bird’s beak at the periphery. This means oxide grows locally over a larger area than what is defined by the mask. This encroachment by field oxide decreases the active device area hence adversely effects the device density. This problem can be solved by poly-buffered LOCOS process.



By locally growing thick layers of oxide, we have achieved Field Oxides 1,2 and 3 for MOS isolation. This local oxidation has taken place by 'LOCOS' technique.

Figure 6.7 Field Oxide Growth by LOCOS technique.

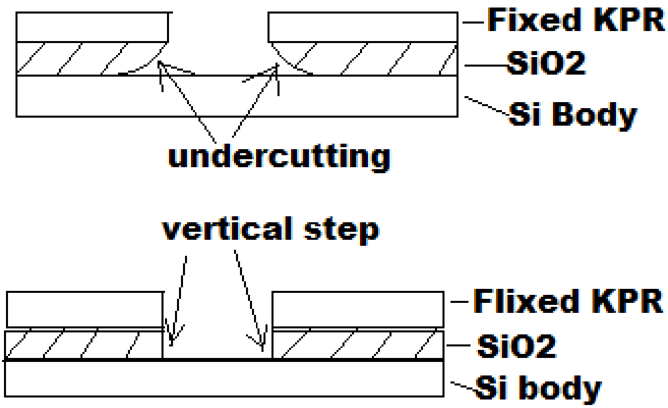


Figure 6.8 Chemical etching is isotropic leading to undercutting whereas Dry etching is anistropic leading to vertical step at the edges.

6.4.5 Poly-buffered LOCOS

In poly-buffered LOCOS instead of SiO2/Si3N4 stack we use a three-layer stack of SiO2/PolySi/Si3N4.

Two stack process uses SiO2/Si3N4 ≡ 40nm/80nm.

Three stack process uses SiO2/PolySi/Si3N4 ≡ 20nm/ 100nm / 200nm.

SiO2 creates a positive stress and Si3N4 creates a negative stress. The two together create zero stress on the Silicon wafer. If the stress was not relieved during LOCOS then defects would be created in Si Wafer leading to low yield of good IC chips.

Thicker nitride and thinner pad oxide in three stack process provides less of a pathway for lateral diffusion of oxidant. Plus poly Si checks the lateral diffusion of oxidant by getting oxidized itself at the edges during LOCOS. This produces a much sharper transition between the oxidized and un-oxidized regions. This allows for tighter design rules and higher device density.

It is obvious the LOCOS produces non-planar interface between Si/SiO2 interface as shown in Figure 6.9. But this non-planar interface gets smoothened out after several runs of oxidation during fabrication. This has been shown in Figure 6.6

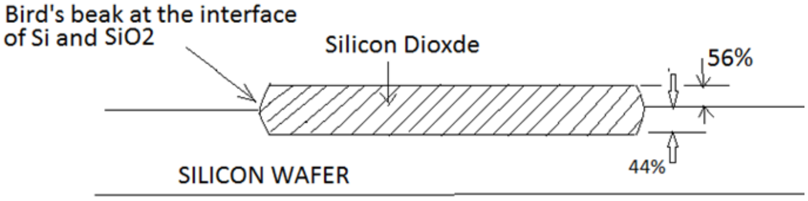


Figure 6.9. Non-planar Si/SiO2 interface as a result of LOCOS. The transition from Si to SiO2 is characterized by Bird’s Beak which has an adverse effect on Device Density.

- Buried Layer Diffusion

In fabrication of MC 1530 the first masked diffusion consists of (n+) type diffusion to provide buried layered to the collectors of the transistors . n+ type region is a heavily doped region which will be subsequently sandwiched between n type epitaxial collector and p type substrate as shown in Fig. (6.10).

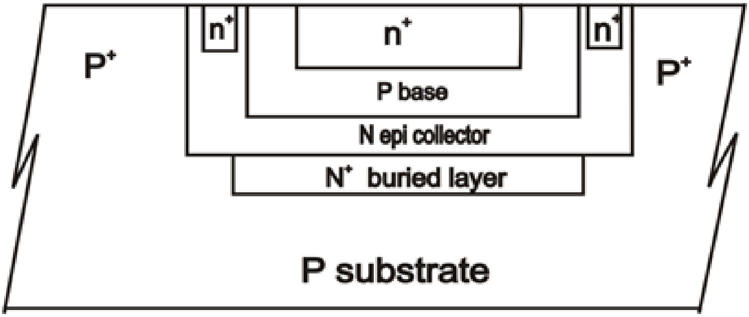


Figure 6.10 Use of a buried N+ layer to reduce collector series resistance.

The buried N+ region has the effect of shunting the high resistance collector body region thereby reducing the series resistance (r_{sc}) of the collector considerably. The reduction of r_{sc} reduces V_{CE} (sat) and increases the driving capabilities of the transistors.

Now as will be seen , after buried layer diffusion there are several heat treatment steps such as epitaxial growth, several oxidation steps and three diffusion steps. In all these proceeding steps, n+ buried layer should not be redistributed or out diffuse and should remain “stay put ” .Because of this requirement of ‘stay put’, Arsenic is used as buried layer dopant as it has a very low diffusion coefficient.

6.4.7_Passivation

Since the oxide layer is impervious to most contaminants that might influence the characteristics of a semiconductor, oxide layer serves as a protection for the semiconductor surface. The oxide layer is said to passivate the semiconductor surface. For this reason the oxide layer is regrown following each diffusion .

Now during masked diffusion impurities diffuse downward as well as laterally under the edge of the oxide as shown in Fig.(6.11) .

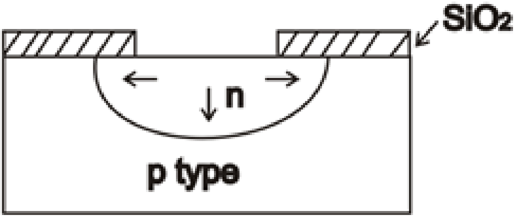


Figure 6.11 Lateral diffusion of impurities under the oxide layer.

Because of lateral diffusion, the p-n junction meets the surface of the chip under the oxide layer hence again passivation prevents device deterioration due to surface leakages .

Thermal oxidation of wafer surface is done in a diffusion furnace as shown in Figure 6.5. A diffusion furnace contains of a central quartz tube where in a 20 inch central zone a flat temperature profile at a given temperature is maintained within ± ½ $_o$ c stability. This permits the batch production of I.C's because hundred of wafers within this 20 inch zone can be simultaneously processed resulting in a tolerance of ±5 to ±10% in diffusion profile. Wafers on a quartz boat are pushed in the central flat zone of the diffusion furnace where temperature is typically maintained between 900$_o$c to 1200$_o$c. An oxidant such as oxygen gas or steam is passed through the central tube at a given temperature for a given time as determined by the thickness of SiO_2. By this process 7000$_o$A thick SiO_2 layer is formed over the wafer. By the use of mask for n+ diffusion for buried layer, and photolithographic technique, suitable windows for buried layer diffusion i.e. for As diffusion are etched out.

Next fabrication step is masked diffusion of 'As' to form buried layers of the transistors used in MC 1530. Solid State Diffusion or selective diffusion of impurities into selected localized surface areas of the wafer is an integral part of I.C. technology. This can be achieved by ion implantation also.

6.4.8 Comments on Solid State Diffusion:

After etching out the SiO_2 mask, the wafers are put in central flat zone of the diffusion furnace shown in Figure 6.14 where temperature is maintained between 800$_{o}$c to 1200$_{o}$c and inert gas (carrier gas) carrying the impurity compound vapour is passed through the furnace. The compound reacts with Si, releasing the dopant. The dopant atoms, dissolved in solid phase of Si, ionize and occupy the substitutional positions in the Si lattice (i.e. the positions of host atoms). The maximum surface concentration of the dopant is No which is determined by the solid solubility of the given dopant into Si.

6.4.8.1 Constant Source Diffusion (pre-deposition)

As long as the partial pressure of impurity compound is maintained at a constant level, the surface concentration is maintained at Ns. The impurity atoms diffuse from the surface to the interior of the wafer resulting in a given impurity profile. The boundary condition in which diffusion is carried out with surface concentration of dopant held constant is known as **constant source diffusion** and this results in complementary error function (erfc) distribution of the impurity in the wafer as shown in Fig.(6.12).

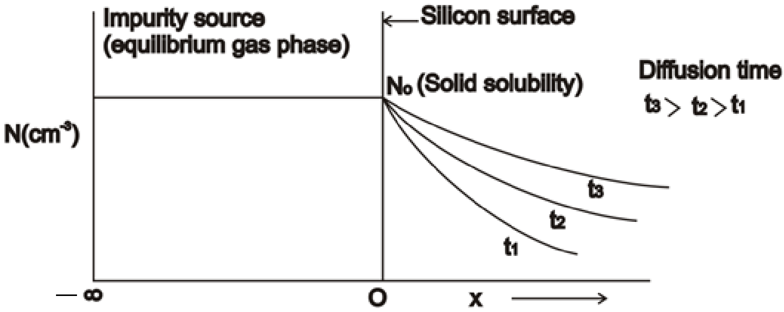


Figure 6.12 Constant Source Diffusion which leads to complementary error function impurity distribution.

(dopant profile) $N(x) = N_s$ erfc

$$\left(\frac{x}{2\sqrt{Dt}}\right)$$

(6.1)

Where D = diffusion constant of the dopant

t = diffusion time

N_s = surface concentration of the impurity.

This constant source deposition is known as pre-deposition. Pre-deposition is carried out with partial pressure of impurity compound high enough so that $N_s = N_o$ (i.e. surface concentration in the wafer corresponds to solid solubility of the impurity in Si). Pre-deposition is carried out in cases where high doping concentration and shallow diffusion is desired.

6.4.8.2 Limited Source Diffusion (Drive-in)

For achieving lighter and deeper diffusion limited source diffusion is resorted to. In limited source diffusion, first through pre-deposition, a layer of impurities are deposited on the surface of the wafer. Suppose total number of impurities deposited is Q. Subsequently in an oxidizing ambient, at high temperature, the final diffusion is carried out. Here the oxidizing ambient forms SiO_2 which seals the impurities in the wafer and fixed quantity of impurities, Q, constitutes the limited source of dopant from which dopants diffuse deeper into the wafer .Hence this is called limited source diffusion or drive in and this results in Gaussian distribution of impurities in the wafer. The Gaussian dopant profile is given by the expression below and shown in Fig.(6.13):

$N(x) = Q/$

$$\sqrt{\pi D t}$$

$\exp($

$$\frac{-x2}{4Dt}$$

$)$ (6.2)

Where Q = total number of impurities on the surface of the wafer at the start of drive in.

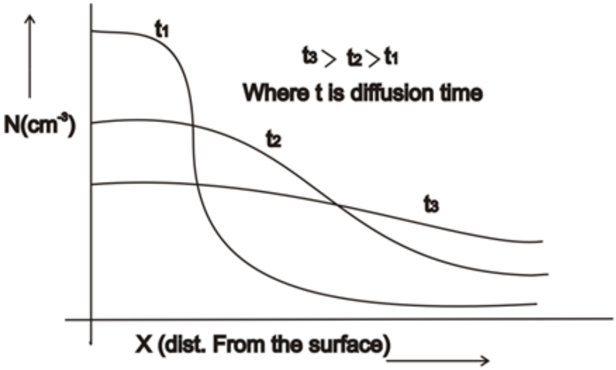


Figure 6.13 Limited Source Diffusion result in Gaussian Distribution. Total area under different curves remain constant with time and Ns decreases with time.

Limited Source Diffusion is also known as Drive in Diffusion. This is done where low surface concentration and deeper diffusion is desired.

The diffusion depth of dopant, which is a fraction of micron to about 10 µm, very critically depends on temperature and time.

Since $D = D_0$ exp

$$\left(\frac{-\Delta H}{RT}\right.$$

$\left.\right)$ (6.3)

D_0, R are constants;

ΔH = activation energy for the diffusion of an impurity,

T = temperature in K.

The critical influence of temperatures on the diffusion profile has forced the semiconductor industry to acquire extremely sophisticated furnaces and controls, capable of holding temperatures in the range of 1000-1300°C to a tolerance of + 0.5°C in order that I.C 's can be manufactured reproducibly. A tolerance of + 0.5°C enables one to achieve a tolerance of + 5 or +10 percent in the diffusion depth which is critically linked to W(base width) a very important device parameter. A still closer tolerance in diffusion depth can be achieved by Ion Implantation.

6.4.8.3 Typical diffusion systems:

An impurity compound in its vapour form, sweeping across the wafers, is used as the corresponding impurity source. Impurity compound reacting with silicon releases the dopant which subsequently diffuses into the wafer. B_2O_3 , BCl_3 and BBr_3 are used as sources of boron (p type dopant) and P_2O_5 , $POCl_3$ are used as sources of phosphorous. In case of solid diffusion source such as P2O5 ,the solid impurity compound is heated in a chamber at a much lower temperature than that of wafers and a carrier gas such as oxygen is passed over the vaporizing $P2O5$, carrying it to the wafers.

$2P_2O_5$ (gas) + 5Si(solid) $\longrightarrow$ 4P + $5SiO_2$(solid).

Through the above chemical reaction , P is released and it diffuses into silicon . Simultaneously oxide layer is formed on wafer surface .A solid source is generally used for pre-deposition. B_2O_3 is used for boron and As_2O_3 for arsenic .

In case of liquid diffusion source , the carrier gas is bubbled though or passed over the liquid impurity compound which is kept in a flask heated to fixed temperature .The liquid vapour is carried by the carrier gas to the wafers where after chemical reaction , the dopant is released. The sources are $POCl3$ for phosphorous and $BBr3$ for boron deposition

A gaseous diffusion source may be used directly or diluted with some other gases.

A typical apparatus for P diffusion using $POCl_3$ is shown in Fig.(6.14).

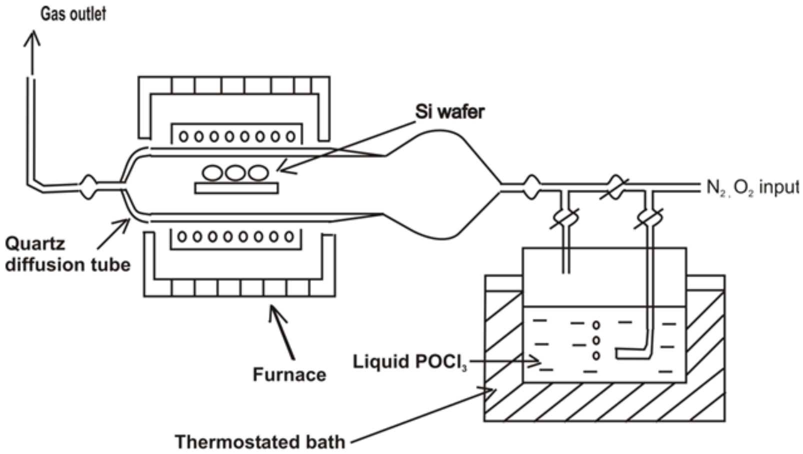


Figure 6.14 Schematic representing typical appratus for P diffusion using POCl3

A carrier gas is bubbled through heated $POCl_3$ (the diffusion source). The impurity compound in vapour form reaches Si wafers and releases P atoms for diffusion. Typically the carrier gas is a mixture of N_2 and O_2 in the ratio of 3:1. This particular ratio leads to the formation of the moderate film of phosphorous –silicon glass necessary for surface protection and yet avoids excessive oxide formation which would otherwise create difficulty in photo resist processing. Practical device structures are limited to not more than three diffusion.

By several diffusion steps the total concentration of impurities below the surface become very large which reduces the electron and hole diffusion coefficients (and hence mobilities) by increasing the impurity scattering and reduces the carrier life time by introducing recombination centers in the crystalline structure. Also the total concentration under surface may reach the upper limit set by solubility so much so that further diffusion and compensation may not be possible. Hence sequential diffusions employed in formation of a device are limited to three.

6.4.8.4 Buried layer diffusion:

Buried layer diffusion is analogous to pre-deposition step. Here solid diffusant source As_2O_3 is used and a high surface concentration of the order of 10^{20}/ccon the Si wafer but shallow diffusion (1 µm) is achieved. Arsenic is chosen as the dopant because it has a diffusion coefficient an order of magnitude lower that of P. Hence buried layer region remains localized during subsequent thermal cycles. Also 'As' has a solid solubility of 10^{21}/cc in Si. Therefore surface concentration of the order of 10^{20}/cc , giving rise to low $R_{sh} = 1$ ohm per square (sheet resistance), is easily achievable.

6.4.9 .Growth of N type epitaxial film over P type substrate:

Fourth important step is the growth of 10 µm thick N type epitaxial film of resistivity 0.5 ohm-cm over a p type substrate of 10 ohm-cm. The epitaxial film has donor concentration of 10^{16} donor impurity atom/cc and substrate has acceptor concentration of 10^{15} acceptor impurity atom/cc. Components are fabricated in epitaxial film hence for isolation from one another and for isolation from the substrate, epitaxial film and substrate are kept of opposite type so that they can be reversed biased and provide isolation.

Epitaxial growth is a process whereby a thin film of single crystal Si is grown from a vapour phase upon on existing single crystal Si wafer with the same crystallographic orientation so that the lattice structure of the resulting layer is an exact extension of the substrate crystal structure.

6.4.9.1 Epitaxial System:

Epitaxial growth system consists of a reaction chamber in which there is a long cylindrical quartz tube surrounded by R.F. induction coil as shown in Figure 6.15. Wafers are put on a graphite boat encased by a quartz sleeve and the boat is pushed in a reaction chamber where localized heating of graphite takes place by RF Induction method. A control console permits the introduction of the wide variety of gases needed to produce epitaxial film with the requisite properties within wide limits.

Silicon tetrachloride with hydrogen is introduced in the reaction chamber. $SiCl_3$ is easy to purify, nontoxic and inexpensive hence utilized generally for epitaxy.

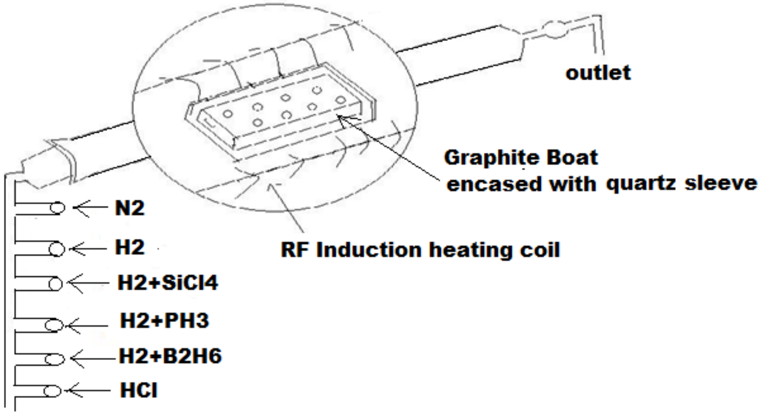


Figure 6.15 A Schemaic Representation of Chemical Vapour Deposition of Si Epi Layer

To achieve the desired impurity concentration in the epitaxial film, Phosphine in correct proportion for N type doping and di-borane in correct proportion for P type, are introduced in $SiCl_4$: H_2 mixture. At the surface of the Si wafer where temperature is almost 1200oc, hydrogen reduction of $SiCl_4$ takes place releasing free Si which skid about on the surface of the growing film until they find a correct position in the lattice and get fastened to the growing film. The reaction, taking place on Si surface only, is known as heterogenous reaction and is given in Eq.(6.4).

1200°C

$$SiCl_4 + 2H_2 \rightarrow Si + 4HCl \; ; (6.4)$$

Equation 6.4 is reversible. In forward direction epitaxial growth of single crystal Si film takes place. In reverse direction Si surface is etched by HCl vapour.

Similar reduction takes place with respect to phosphine or di-boranes. This results in the incorporation of phosphorous or boron atoms in the growing silicon lattice. By this process a variety of controlled resistivity and complex impurity profiles can be achieved. There are certain precautions which are imperative for epitaxial growth. No matter how well the wafers are cleaned, there are mechanical imperfections, silicon dioxide, residual dust and other contaminants on the surface of the wafer. These surface defects nucleate crystal imperfections which manifest themselves as bumps, pits and stacking faults. A scratch on the substrate would show up in the film also. All these faults and imperfections affect the reproducibility and yield of I.C. chips. (Yield means percentage of I.C. chips which has normal performance). It has been found that if vapor phase HCl is passed over the wafers before epitaxial growth, to etch away a few microns at the surface of the wafers, then scratches, dust and contaminants are removed to a great extent, resulting in epitaxial films which are completely free of bumps and pits and where stacking faults are reduced several orders of magnitude.

Epitaxy process steps in manufacture of MC 1530.

1. SiO_2 layer, which was used as a mask for buried layer diffusion, is completely etched away using HF (hydrofluoric acid) as an etchant.

2. The cleared wafers are placed on the boat and moved into the reaction chamber.
3. The system is closed and N_2 admitted to flush out the air.
4. H_2 is passed through the chamber at the rate of 30 liters/min.
5. RF heating is turned on and substrate wafers are maintained at 1200°C ± 2°C.
6. When the temperatures have been stabilized, HCl is admitted. H_2: HCl $\equiv$ 100: 1. HCl etches away 3 μm of the substrate surface in 6 minutes.
7. At this point HCl turned off and $SiCl_2$ and phosphine turned on.

H_2: $SiCl_4 \equiv$ 800:1 and H_2: phosphine $\equiv 5 \times 10^8$: 1. This reaction continues for 50 minutes.

1. Phosphine and $SiCl_4$ are turned off, RF heating is turned off and wafers cooled in pure hydrogen.

These steps result in 26 μm N type epi-film of 0.5 ohm-cm. resistivity having 10^{16} Phosphorous atoms per cc. (This process yields material in which thickness and resistivity are controlled to within ±5 percent).

0.5 ohm-cm resistivity of epitaxial film is an ideal compromise between high resistivity, required for high BV_{CBO} and low C_{jc} and high transit frequency.

The transistor to be manufactured in this epi layer will have large BV_{CBO} and wide frequency response.

6.4.10 Isolation Diffusion

After the epitaxial growth cycle, SiO_2 is grown over the epitaxial layer by standard oxidation process. SiO_2 film is about 5000 A° thick or 500nm. By using Mask No. 2 and photolithography process, desired windows are etched out in SiO_2 film and the wafer is ready for "isolation diffusion".

The wafer is put in diffusion furnace and Boron (P-type) dopant atoms are diffused into the silicon wafer as described before in diffusion section. A long 'drive-in' is required so that B ions diffuse through the epi layer to the P type substrate. The areas that remained covered with SiO_2 are now isolated islands of N-type epitaxial Si surrounded by P-type Isolation boundaries. if reverse biased n-p junction formed around each n region and between any two n regions these junctions result in back to back diodes out of which at least one is always reverse biased no matter what polarity voltage appear between the two islands. Since B_2O_3 (the impurity compound) is used as the diffusant source so invariably a new layer of SiO_2 grows over the diffused P-type region and pre-existing oxide over the N-epi-regions grows thicker as is evident from Figure (6.16).

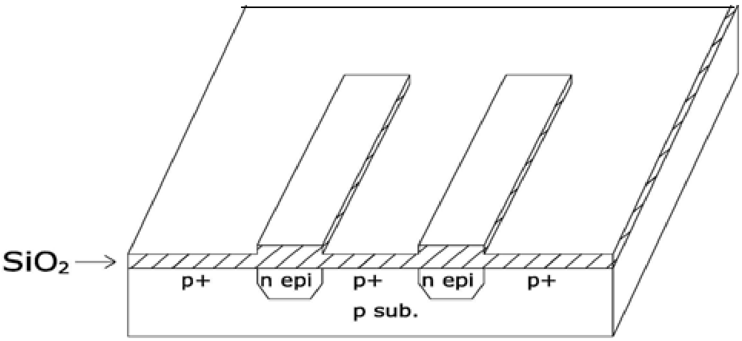


Figure 6.16 The Wafer cross-sectional area after isolation diffusion.

6.4.11 P type diffusion for Base and Resistors

Using photo Mask No. (3) and photolithographic process, windows are etched in the existing SiO_2 layer. Boron (P-type) atoms are “driven in” to a depth of 2.7µm (x_j) into the N-type isolated islands for forming the Base and Resistors. Here surface concentration is $N_s = 5 \times 10^{18} (cm^{-3})$. This corresponds to sheet resistivity (R_s) 200 ohms/cm^2 where $R_s = \rho / x_j$. In the process of the diffusion , a new layer of SiO_2 is again grown over the newly diffused P-type regions and pre-existing layer become thicker .

6.4.12 N+ diffusion for Emitter and Collector Contact

Using photo Mask No. (4), by photolithographic process oxide coating is selectively etched to provide windows for Phosphorous (N-type) diffusion. Windows are etched in Emitter regions of the transistors and in Contact areas of the Collector. Why N+ diffusion is provided as Collector Contact areas? Aluminum (Al) is going to be used for making ohmic contacts to the terminals of the device and also for interconnections. Al is P type materials. Hence Al makes ohmic contact with P type region but makes rectifying contacts with N type region. This rectifying contact is also known as Metal-Semiconductor Schottky Diode. N type region should be made degenerate (semi-metal) in order to make ohmic contact with Al. Therefore N+ diffusion is provided in the contact areas of the collector which is lightly N type of the order of 10^{16} Phosphorous atoms in Si epitaxial film.

Phosphorous (N type) is used for pre-deposition (constant source diffusion) into the Si wafer. Surface concentration of $N_s = 10^{21}(cm^{-3})$ is maintained and diffusion depth of $x_j = 2\mu m$ is achieved in constant source diffusion. This results in $R_s = 2.2\Omega$/square = $2.2\Omega/\square$. At this point the junction formation in the monolithic circuit is complete. In the process of diffusion a new layer of SiO_2 is formed all over the wafer.

The diffusion profile of the transistor of the monolithic circuit is given in Fig.(6.17).

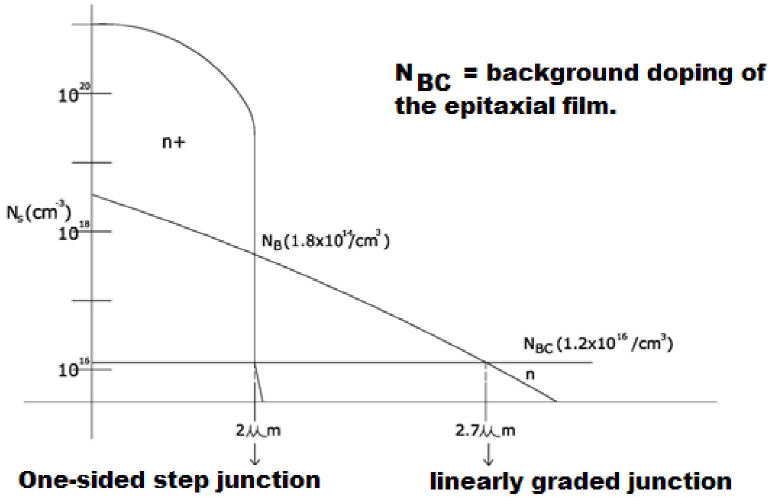


Figure 6.17 The absolute dopent profile of a vertical IC NPN transistor.

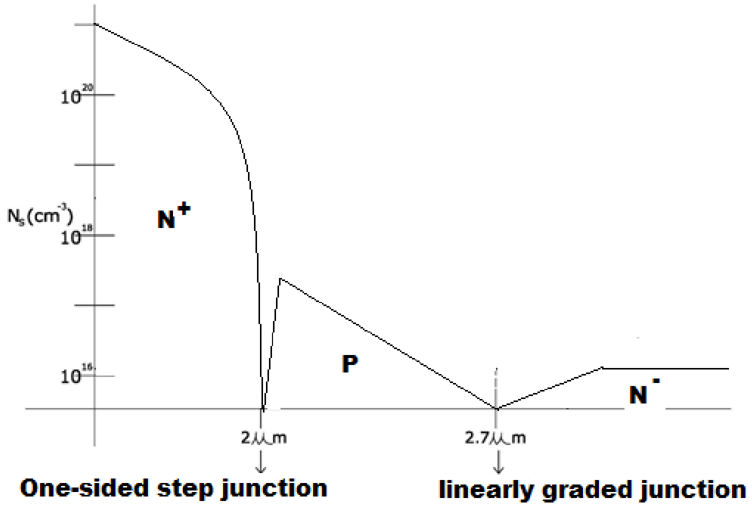


Figure 6.18 Dopent concentration profile after compensation along the longtitudinal axis.

Looking at the diffusion profile it is clear that E-B junction is one-sided abrupt (or step) junction and

B-C junction is linearly graded junction.

6.4.13 Interconnection of the various components of the monolithic circuit.

By using mask No.5 a new set of windows are etched out in the regions of contact pads to the components.

By vacuum deposition method a thin film of Al(thickness 1000Angstrom) is deposited evenly over the entire surface of the wafer. Then by use of mask no. 6 the desired interconnection pattern between the components are formed by photolithographic processes. The undesired Al is simply etched out by the use of suitable etchant(namelt NaOH or KOH).

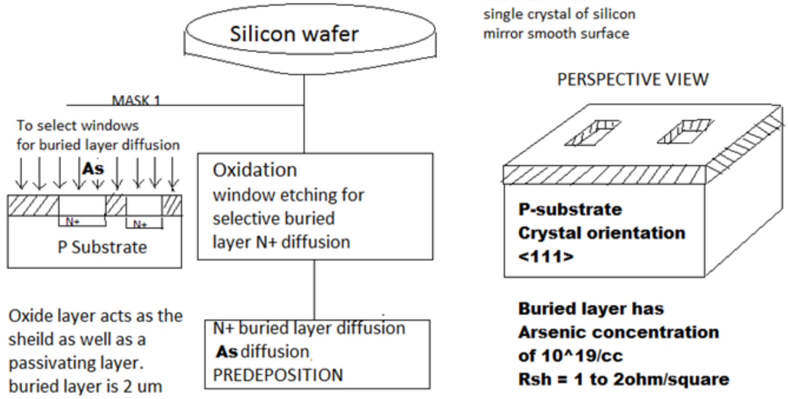
Next the wafers are scribed. Individual dies or chips are mounted on headers and wire from contact pads of chip are bonded to leads of the headers and hermetically sealed either as flat package or T0-5 package. All these steps have been previously described.

Thus MC-1530 is ready for customer’s use.

6.4.14 Flow-Chart of fabrication of MC-1530

The sequence of fabrication steps taken in the fabrication of MC-1530 are given in Fig.(6.19)

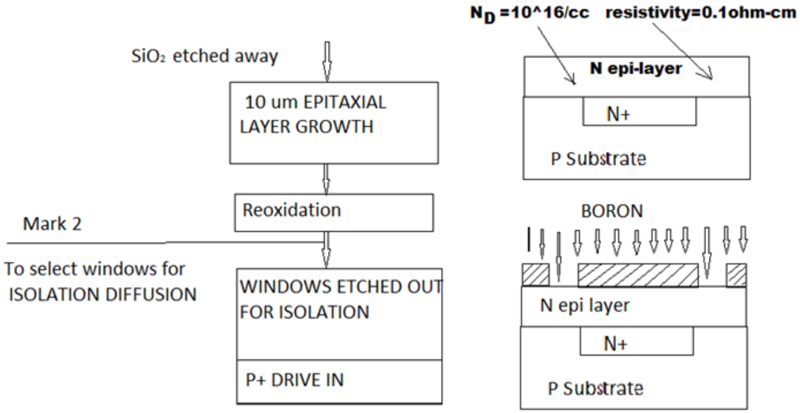
FABRICATION STEPS OF AN IC CHIP

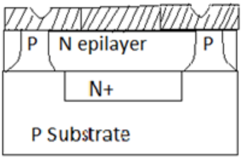


diameter = 200mm
thickness t=400um
single crystal of silicon
mirror smooth surface

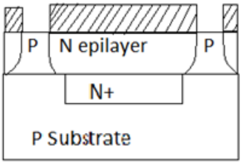
Oxide layer acts as the sheild as well as a passivating layer.
buried layer is 2 um thick

Buried layer has
Arsenic concentration
of 10^19/cc
Rsh = 1 to 2ohm/square





ISOLATION P+ DRIVE IN

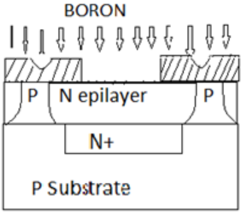


REOXIDATION

MASK 3

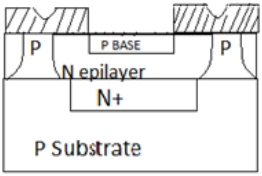
To select windows
for P type
Base Drive-in

WINDOW ETCHED OUT FOR P type BASE drive in

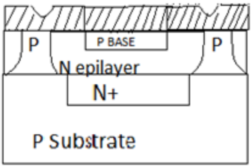


Doping Density of base
NA =10^17/cc
Rsh =200ohm/sqr
Diffusion depth=3um

BORON P type BASE drive - in



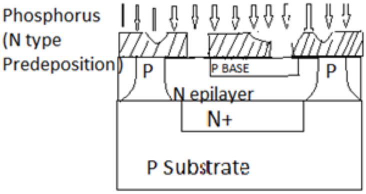
REOXIDATION AFTER P type Base Diffusion



MASK 4

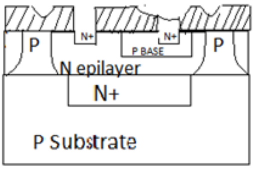
To select windows for
Emitter and Collector
contact diffusion

WINDOWS FOR EMITTER AND COLLECTOR CONTACTS

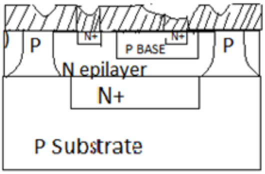


ND=10^19/cc
Rsh=1 to 2 ohm/sqr
Diff.depth= 2micron

N+ predeposition
for EMITTER and
COLLECTOR CONTACTS



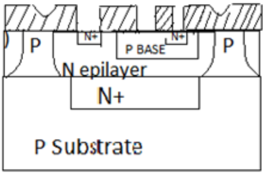
Reoxidation after
N+ predeposition



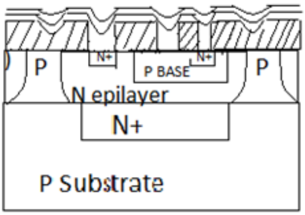
MASK 5

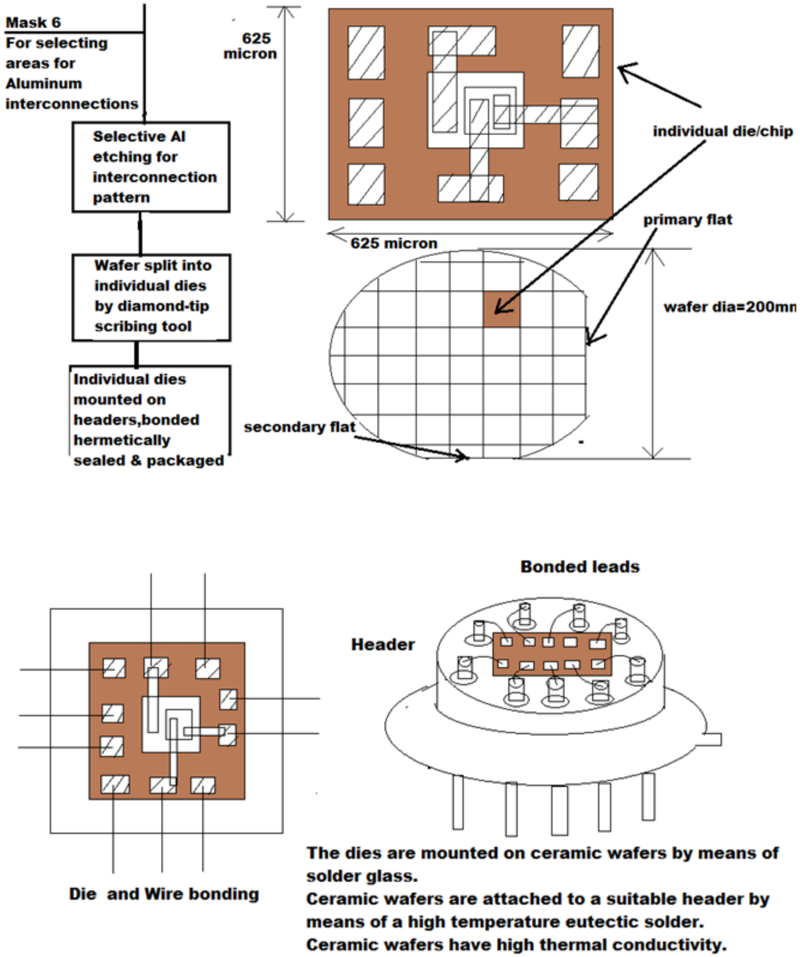
To select windows
for Al contact
pads of Emitter,
Base,Collector

WINDOWS FOR
Al CONTACTS



Aluminization of
upper Surface
in Vaccum
Evaporation Chamber




Mask 6
For selecting areas for Aluminum interconnections
Selective Al etching for interconnection pattern
Wafer split into individual dies by diamond-tip scribing tool
Individual dies mounted on headers,bonded hermetically sealed & packaged
625 micron
625 micron
individual die/chip
primary flat
wafer dia=200mn
secondary flat
Bonded leads
Header
Die and Wire bonding

The dies are mounted on ceramic wafers by means of solder glass.
Ceramic wafers are attached to a suitable header by means of a high temperature eutectic solder.
Ceramic wafers have high thermal conductivity.

Figure 6.19 Flow-Chart of MC-1530 fabrication steps.

Chapter 6. IC Fabrication Technology from 60s to date_Part 2

Part 2 of Chapter 6 describes the fabrication technology of vertical NPN,IC diode, IC resistance, IC capacitance, Lateral PNP, breakdown voltages of NPN and PNP transistors and what are the layout considerations in IC fabrication.

Chapter 6_IC Fabrication Technology from 1960 todate_Part 2

6.5 Vertical NPN Transistor realization in IC Technolgy

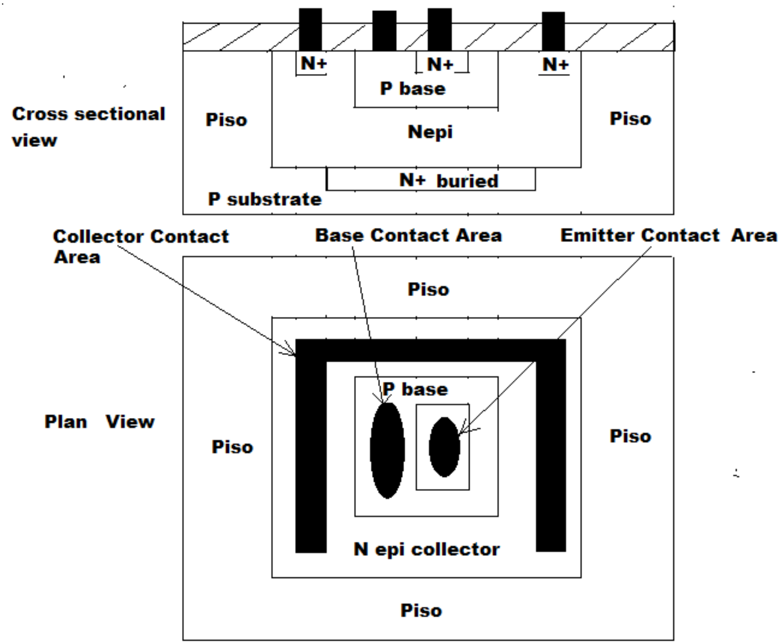


Figure 6.20 Stripe Geometry Vertical NPN

Transistor in IC chip.

Section 6.6 Diode realization in IC Technology

Diode realization: A ‘n - p – n’ transistor is essentially a back to back diode. Hence any of the two junction diodes or both may be used for diode application. Therefore there are five possible combinations of e, b, c terminals by which diode can be achieved. But out of this c, b terminals shorted and used as anode whereas emitter terminal used as cathode gives a diode which has optimum performance parameters. The configuration is given in Fig.(6.21).

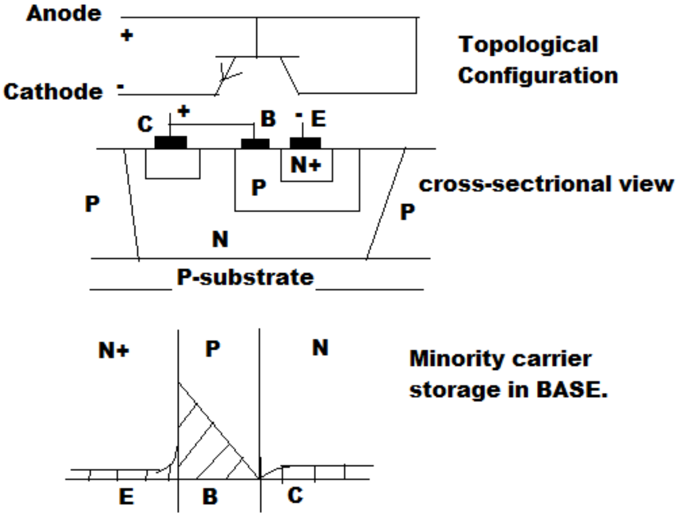


Figure 6.21 Transistor connected in Diode Configuration. This is one possible configuration.

c-b shorted diode configuration of transistor has minimum storage delay or recovery delay time t_d.

(t_d = delay involved in switching a diode from ON to OFF conditions). t_d is typically 9nsec. This diode is still a transistor in active region, since c-b is shorted hence it is still reverse biased. Diode isolation island is also provided with a buried layer so as to reduce series collector resistances. This enables the diode to give least forward bias voltage. Also this has the least leakage current flowing under reverse bias condition since the leakage current of E-B junction only comes into picture (E-B junction has smaller cross sectional area and has least depletion width and leakage current I_O in reverse bias condition is predominantly the generation current in depletion width namely $I_g = q\ g\ X_m\ A$

g = generation rate,

X_m = depletion width at the given reverse voltage,

A = cross sectional area of the junction.

Section 6.7 Resistance realization in I.C. Technology:

p type diffusion for base gives a sheet resistance R_{sh} = 200 ohms per square. By taking long strip like configuration of base diffusion, various resistances are realized. In Fig. (6.22) the resistance is given.

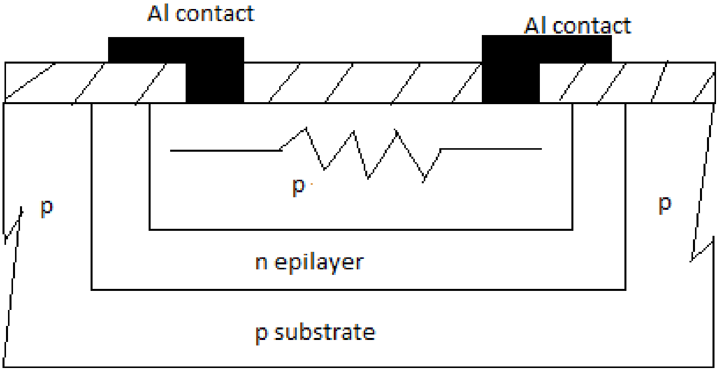


Figure 6.22.A. Cross Sectional View of a typical I.C. resistance

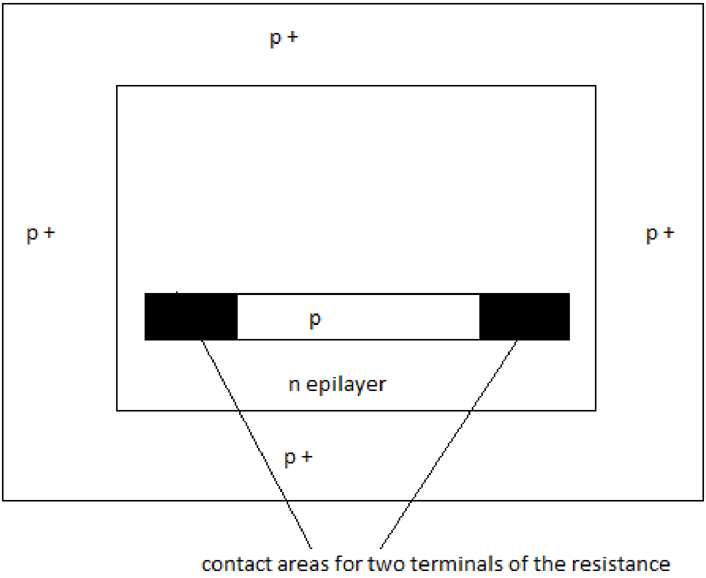


Figure 6.22.B. Plan View of a typical I.C. resistances.

Section 6.8 Capacitor realization in IC Technology:

In uA 741 the frequency compensating capacitor C_f = 30 pF is monolithically implemented. For I.C. realization of C_f , silicon dioxide is used as the dielectric sandwiched between metallic Al layer and n+ degenerate layer. The n+ region is deposited during emitter diffusion step. The cross sectional view of the capacitor is given in Fig. (6.23)

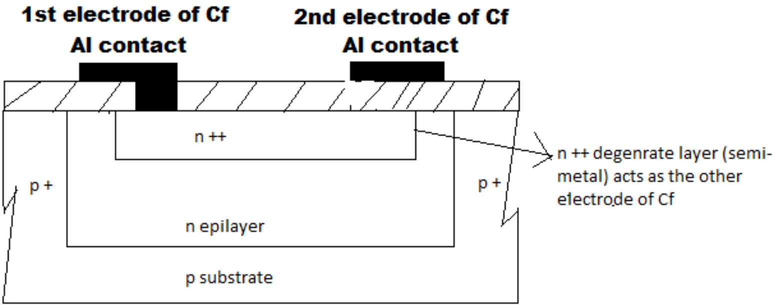


Figure 6.23 Cross-sectional view of the frequency compensating capacitor which introduces a dominant pole at 10Hz in uA741.

“Introduction to Integrated circuits “ by Grinich and Jackson, the photomicrograph of uA 741 is given. As can be seen from the photomicrograph, 30 pF capacitor takes a significant chip area. Therefore it is neither economical nor practical to integrate circuits using capacitances.

Section 6.9 p-n-p transistor realization:

Here the n epilayer is used as the base and p diffusion for forming the base of n-p-n transistor is used for forming the collector and emitter of p-n-p transistor. The cross-sectional view of p-n-p transistor which is also known as lateral transistor is shown in Fig. (6.24). This is known as lateral transistor as the minority carriers flow laterally.

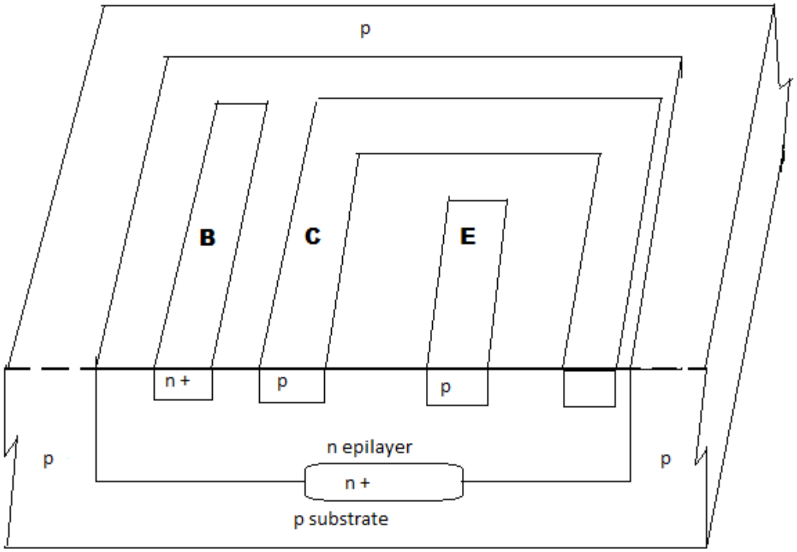


Figure 6.24 Perspective view of Lateral PNP transistor.

As is evident from Fig. (6.24) W, the base width, is limited by the resolution of the photolithographic technique which is 10um in 1960s. Also punch through has to be avoided. (Punch through occurs when C-B depletion region overlaps E-B depletion region; at this point excessive current flows and transistor action is lost). Therefore in lateral transistors W is 5 to 10um. This large W causes a drastic deterioration of short circuit current (β_{fo}) and transit frequency (ω_T)

From our device physics we know that Transit Time

$= \tau_{transit} = W^2/(2D_B)$ and $\beta_{fo} = (\tau_{recombination}/\tau_{transit})$

Also since P doping of emitter (10^{17}/cc) and collector (10^{17}/cc) is heavier than N doping(10^{16}/cc) of the epilayer which is base here, the depletion width extends into base region predominantly. This causes $\delta W/\delta V_{CB}$ to be significant which leads to significant Early Effect hence h_{ob} and h_{oe} (or g_c and g_o) are large (the slope of I-V characteristics is large) But here $BV_{EBO} = BV_{CBO}$ high is quite high.

Because of poor performance parameters of lateral transistors, the latter are fully avoided for high frequency applications. But in low frequency applications (typically below 1 MHz) a mix of n-p-n and lateral p-n-p can give very favorable performance characteristics. Therefore such a mix is used in uA 741.

Section 6.10 Breakdown voltages of n-p-n transistor in MC1530

Referring to the doping profile of n-p-n transistor in Fig. (6.17) it is seen that

background doping of the epi-layer is $N_{BC} = 1.2 \times 10^{16}$ (cm^{-3})

and $N'_B = 1.8 \times 10^{17}$ (cm^{-3}), emitter doping is $N_E = 1.0 \text{ X } 10^{20}$ (cm^{-3}) and W = 0.7 um.

Therefore

BV_{CBO} (determined by N_{BC}) = 55V.

BV_{EBO}(determined by N_{B}’) = 7V.

$BV_{CEO} = BV_{CBO}/ (\beta_{o})_{1/MillerIndex} = 23V$

[Substrate pnp transistor:- Here p substrate is used as collector and n epilayer is used as base and p layer, for the base diffusion in npn transistor, is used as emitter. Since here W depends upon diffusion depths therefore much better performance parameter as compared to lateral pnp can be achieved. But substrate pnp can be used only in CC configuration.]

Doping profile of pnp transistor is given in Fig. (6.25)

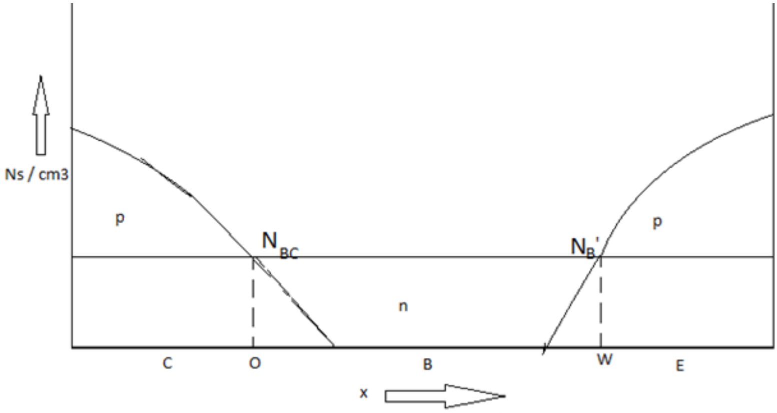


Figure 6.25 Doping profile of lateral pnp transistor

As is evident from Fig. (6.25)

$N_{B}' = N_{BC}$

Therefore $BV_{CBO} = BV_{EBO} = 55v$.

Therefore if lateral pnp transistor is used in the input circuit of Op-Amp (as it is done in uA741) a much larger V_{id} (differential mode signal) can be handled by the Op-Amp.

Section 6.11 Design Considerations in the circuit layout of MC1530/31.

We have already seen the fabrication steps involved in the manufacture of MC 1530 or MC 1531. Now we will examine the design considerations involved in the circuit layout of MC -1530 on the silicon chip. MC-1530 and MC-1531 are identical circuits except for the input differential stage. MC-1530 is a direct input to a differential stage with $R_{id} = 10K$ whereas MC-1531 is a darlington pair input to the differential stage with $R_{id} = 1M$.

We have already seen that N epi layer of the isolated island forms the collector of the transistor or the anode of the diode. (We have shown in Sec. 6.5 that transistor is used in diode configuration by shorting collector to base). Therefore N epi layer of any isolated island will always be at some positive potential as dictated by d.c. quiescent condition of the circuit. These positive potentials set the proper operating condition of the circuit as well as they help isolate the components from the substrate and from one another.

Refering to the circuit in Fig. (6.26) we see that there are eight different collector voltages and four different Diode anode voltages:

1) input darlington pair Q_1 and Q_2 correspond to one collector voltage,

2) input darlington pair Q_4 and Q_5 correspond to second collector voltage,

3) Q_6, Q_8 and Q_{12} correspond to the third voltage namely Vcc collector voltage since they are in CC (Common Collector) mode,

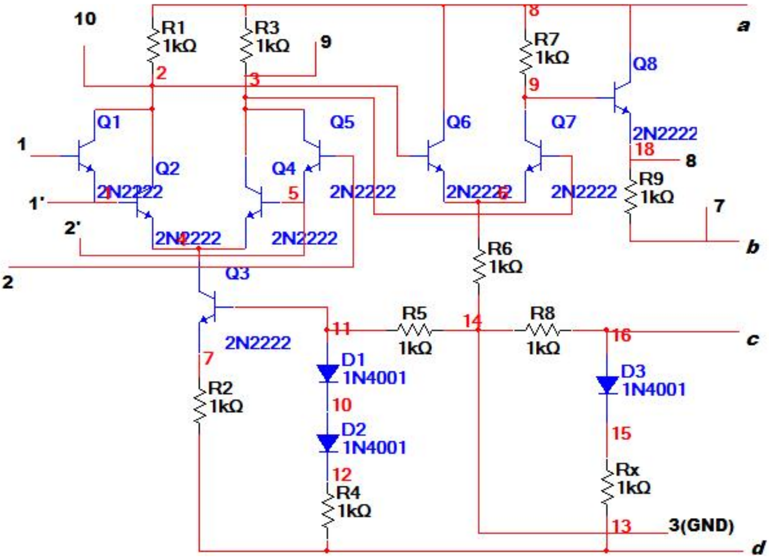
4) Q_3, Q_7, Q_{10}, Q_{11}, Q_{13} correspond to fourth, fifth, sixth, seventh, and eighth collector voltages;

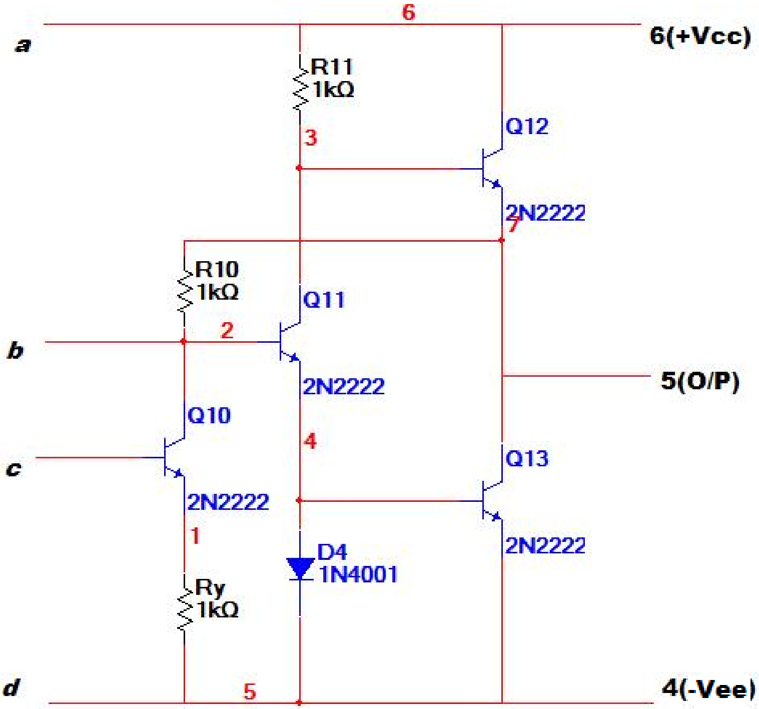
5) there are four different diode anode voltages corresponding to D_1, D_2, D_3, D_4.

Corresponding to these twelve different quiescent voltages which are Vcc or less but always greater than $-V_{EE}$, we have twelve isolated islands as seen in Fig. (6.27) where each isolated island is marked by thick border lines.

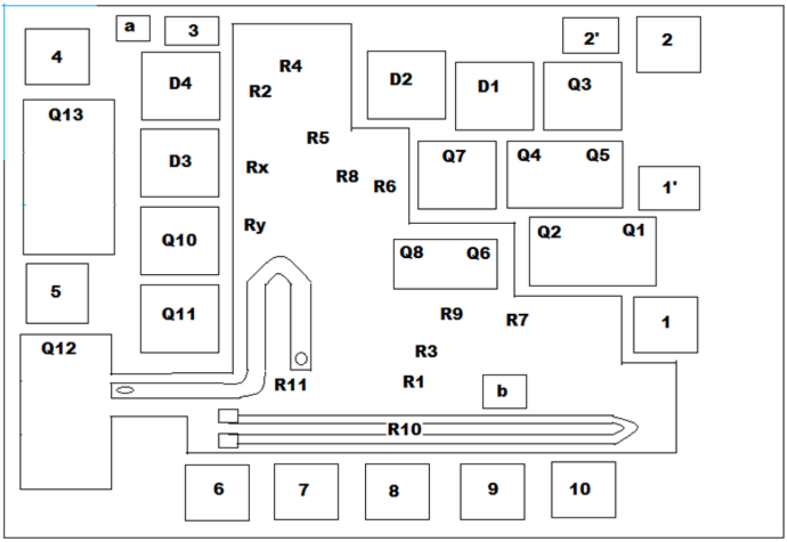
The isolated islands containing Q_6, Q_8 and Q_{12} and connected to Vcc contains all resistances R_1 to R_{11}

and Rx and Ry. Resistances are realized by an elongated strip of p diffusion and any point on the resistances is less than Vcc or at most Vcc. In order that resistances may be isolated from n epi layer and hence from one another, n epi layer i.e. the isolated island must be at most positive potential i.e. Vcc. Then only all resistances will be reverse biased with respect to the n epi layer.





Fig(6.26) Complete circuit diagram of MC 1530/31 chip, with the points of different contact pads marked



Fig(6.27) Twelve isolated islands and placement of transistors, diodes and resistors on the islands have been indicated. The detailed structure of the devices have not been given.

In Fig. (6.27) complete layout of MC-1530 is given. In the chip layout following are the contacts connected to the header leads:

Contact(a) connects substrate to contact(4) which is held at -VEE.________________Contact(b) connects the isolated island containing resistances and CC transistors to contact(6) which is held at +Vcc.____

and (2') are input connections for MC-1530,

__________________________________Contact(1) and (2) are input connections for MC-1531____Conta

is
output,
is
ground,
, (8), (9)and (10) are monitoring points at V_{B11}, V_{E8}, V_{C5}, V_{C1} respectively.

In MC 1530-31, each transistor and diode configuration of transistor contains buried layers. The epilayer is 0.5 Ω cm. This allows BV_{CEO} to be 20V. But because of buried layer the collector series resistance (r_{sc}) or saturation resistance are less than 50Ω. Hence with $V_{CC} \pm 9V$, an output swing of $16V_{p\text{-}p}$ is achieved without endangering the output transistors Q_{13}, Q_{12} and at the same time r_{sc} does not affect the ability to drive loads down to 500Ω.

In the design of MC1530-31, the critical d.c. levels are kept a function of resistor ratios and not of absolute values. This is because in I.C. technology, diffused resistance ratios with much closer tolerance (typically 5 to 6 %) can be achieved as compared to the tolerance of absolute values (30 to 50 %). That is matching tolerance is much narrower than absolute tolerance. Therefore output d.c. level (V_0) is a function of only resistor ratios – not of the absolute value of any resistor or of the beta of any device of the circuit. V_o is given by Eq.(6.5)

$$V_o = V_{EE}\left(\frac{R_{10}}{R_8} - \frac{R_{10}}{R_9} - 1\right) - V_E\, Q_8 \left(\frac{R_{10}}{R_9}\right) + 1.4\ (1 + \frac{R_{10}}{R_9} - \frac{R_{10}}{R_8})$$

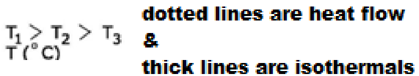
= 0V -------------------------(6.5)

Since the resistors are more or less temperature tracked therefore ratios remain temperature independent. Hence output voltage V_o does not vary with temperature. Only mismatch in the input circuit (with respect to beta, V_{BE}, I_{CO}) causes a finite V_o(i.e. offset voltage), and with temperature variation this mismatch gates magnified and causes a drift of V_0

In Fig 6.26 it is seen that output transistor Q_{12} and Q_{13} have three times cross-sectional area as that of the emitter base junction of D4. Hence output transistors carry 3 times as much current as D4. From dc bias condition,

$I_{Q12} = I_{Q13} = 3(I_{D4}) = 308mA;$

Because of this enhanced signal handling capacity MC1530 will drive load resistors as low as 1k to a maximum p-p swing of 11 volts with Vcc = ±6 volts. Also because of high quiescent current in output stage maximum power is dissipated in Q12, Q13. Most of the dissipated energy is sunk by the header but not without heating the chip. Hence during the operation invariably a temperature gradient and a set of isothermals are set up on the chip as shown in Fig (6.28).



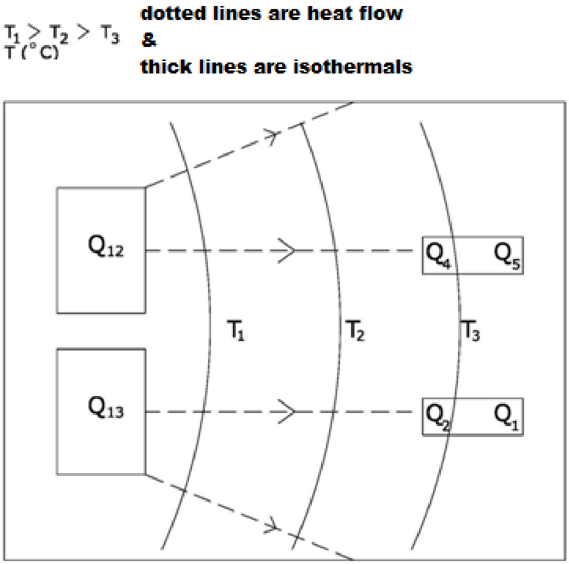


Figure 6.28 Patterns of isothermals on MC1530/31 chip.

As can be seen in Fig.(6.28) input pairs Q2 and Q4, Q1 and Q5 are kept on the least isothermal, farthest away from the output transistors, and in close vicinity to each other. The result is that there is least heating of the input pairs because of high output power dissipation, maximum temperature tracking of the input pairs and diffusion profiles of the input pairs very closely match to one another. Because of close proximity and the resultant diffusion profile matching, V_{BE} and betas of the input pairs of transistors very closely match hence very low input offset voltages and currents are caused typically $V_{OS} = 2mV$. Because of excellent temperature tracking of the input pairs, the long term drift of temperature appears mainly as common mode signal thereby minimizing the long term drift of V_{os} to 3.8μV/°C. Both these parameters V_{os} and low drift are indicative of proper layout of the devices on the chip.

As can be seen in Fig.(6.27) D_1, D_2 are in close proximity to Q_3 and the three are more or less on the same isothermal, D_3 is in close proximity to Q_{10} as well as the two are on same isothermal, and D_4 is in close proximity to Q_{13}. Each of these are bias current sources. Because of matching of the characteristics and because of temperature tracking of the devices in the current sources, each current source is the current mirror of the reference current. And because of the use of diode configuration of the transistor which results in heavy voltage shunt feedback the reference current are highly stabilised against temperature changes. This is also known as diode compensation. Hence almost drift free near ideal current sources are achieved. Because of the stability the bias sources, the ac performance of this chip is highly stabilised. The open loop gain A_d = 72dB ± 1.5dB from T = -55°C to T= +125°C.

The monolithic IC chip MC 1530-31 is mounted on 10 pins TO-5 packages. Two battery biases V_{cc},

$-V_{EE}$, output and ground leads require four pins, two pins are used for inputs and four pins are for internal connections.

SSPD_Chapter 6_Part 3_The challenges of IC Technology

SSPD_Chapter 6_Part 3 describes the challenges faced by IC fabrication technology in course of increased micro miniaturization.

SSPD_Chapter 6_IC Fabrication Technology from 1960s to date (2011)_Part3.

The tortuous path and the challenges in the development of IC Fabrication Technology.

Gordon Moore predicted that number of transistors on integrated circuits (a rough measure of computer processing power) will double every 18 months at a minimum cost. It became a self fulfilling prophecy. *Moore's Law has become a yardstick of our progress as we harness the cunning of NATURE's design strategies.*

There was a constant pressure on Chip manufacturers to integrate functionally complex chips. To meet this target it became a necessity to follow Moore's Law of micro-miniaturization. Following this law through aggressive vertical and lateral scaling rapid advances took place and major milestones were established in the field of IC Technology as already mentioned in the Road Map of Computer Development.

In 1964 the third generation Computer using IC came into market. This was known as Table-Top

Computer PDP-8 . This was compact. There was much less power requirement and much more cost affordable.

In 1971 the first microprocessor chip (μP4004) was marketed by INTEL . This laid the foundation of Desk-Top computers belonging to Fourth Generation Computers.

μP4004 was followed by 8085, 8086, 80386, 80486, Pentium I, Pentium II, Pentium III and Pentium IV. Today in 2009 such powerful quad processors have been commercialized that Nvidia advertises its new workstation, the Tesla, as a “personal supercomputer”. It clusters 4 Nvidia C1060 processing boards, each of which unites 240 graphic cores to process instructions at nearly teraflop speed.

Table 6.6. Transistor Junction Depths and Doping Changes with years. [Muller & Kamini, Solid State Devices; “Modelling of SiGe heterojunction transistors:200GHz frequencies with symmetrical delay times”, Jochen Eberhardt & Erich Kasper, Solid State Electronics, 45(2001), 2097-2100.] .

	Amplifying(Junction Isolated)1964	Amplifying(Junction Isolated)1980	Switching(Junction Isolated)1980	Switching(Junction Isolated)1980	Amplifying(oxide Isolated)2001H
Surface area	212.5μm × 250μm				20 μm × 1μm
Epi Film Thickness_Resistivity of epi layer	10 μm_0.5Ω-cm	10μm_1Ω	3μm_0.3-0.8 Ω-cm	1.2μm_0.3-0.8 Ω-cm	370nm W_C = 50nm
Buried Layer Sheet R_{Sheet} Updiffusion	1Ω/□_2.5 μm	20Ω/□_2.5μm	20Ω/□_1.4μm	30 Ω/□_0.3 μm	0.7 Ω/□_0.2 μm
Emitter Diffusion Depth_Sheet Resistance of Emitter	2.5 μm_1Ω/□	2.5μm_5Ω/□	0.8μm_12Ω/□	0.25μm_30Ω/□	W_E = 110nm_378 Ω/□_Emitter contact_5.2 Ω/□
Base Diffusion Depth_R_{SHEET}	3.5 μm_200Ω/□	3.25μm_100 Ω/□	1.0μm_200 Ω/□	0.5μm_600 Ω/□	122nm260 Ω/□
Base Width	1 μm	1 μm	0.5 μm	0.25 μm	W_B = 12nm
Substrate Resistivity Crystal Orientation	10 Ω-cm <111>	10 Ω-cm <111>	10 Ω-cm <111>	5 Ω-cm <111>	

Table 6.7. Dimension Scaling in MOSFET over the last decade.

MOS	1967	1997	1999	2001	2003	2006
L(µm)	10	0.25	0.18	0.13	0.1	0.07
DRAM(Gbit/cm2)	64M	0.18	0.38	0.42	0.91	1.85
Junction Depth(xj)nm	1000	100	70	60	52	40
Interconnection Pitch(nm)	2000	600	500	350	245	130

Table 6.6 and Table 6.7 shows the strides taken in microminiaturization. This has meant a lot of energy, resources and time invested in technology improvement to achieve this kind of scaling.

As of 2004, current PC processors are fabricated at the 130 nm and 90 nm levels, with 65 nm chips being announced by the end of 2005. A decade ago, chips were built at a 500 nm level. Companies are working on using nanotechnology to solve the complex engineering problems involved in producing chips at the 45 nm, 30 nm, and even smaller levels—a process that will postpone the industry meeting the limits of Moore's Law.

Moore's Law of Integrated Circuits was not the first, but the fifth paradigm to provide accelerating price-performance. Computing devices have been consistently multiplying in power (per unit of time) from the mechanical calculating devices used in the 1890 U.S. Census, to Turing's relay-based "Robinson" machine that cracked the Nazi enigma code, to the CBS vacuum tube computer that predicted the election of Eisenhower, to the transistor-based machines used in the first space launches, to the integrated-circuit-based personal computers. In the process of miniaturization we have met with seemingly unsurmountable obstacles which in due course of time with new technological breakthroughs and innovations have become surmountable.

These technological breakthroughs and innovations have meant that R&D, manufacturing, and test costs have increased steadily with each new generation of chips. As the cost of semiconductor equipment is expected to continue increasing, manufacturers must sell larger and larger quantities of chips to remain profitable. (The cost to "tapeout" a chip at 0.18um was roughly $300,000 USD. The cost to "tapeout" a chip at 90nm exceeds $750,000 USD, and the cost is expected to exceed $1.0M USD for 65nm.)

"What we didn't realize then was that the integrated circuit would reduce the cost of electronic functions by a factor of a million to one," Kilby said.

Placing several million transistors on a piece of silicon the size of a fingertip is intricate and exacting. Precision associated with chip manufacturing is measured in microns and increasingly in fractions of microns. A micron is one-millionth of a meter, or about one onehundredth of the diameter of a human hair. Maintaining this level of precision demands chip production environments that are 1,000 times cleaner than today's cleanest surgical operating rooms.

Challenge number 1: “Clean Rooms for IC fabrication”.

Challenge number 2: “Patterning or Lithography”.

Challenge number 3:”Transistor Options”. Classical scaling ended in 2001.

Challenge number 4:”Interconnect options”.

Challenge number 5:”High density memory beyond SRAM is required in today’s devices.”

Challenge number 6:”System Integration”.

Solving these problems INTEL is moving to 32nm technology.

1960s was dominated by Bipolar Technology and TTL Logic Gates operating at 5V.

The earliest challenge of 1960s was the production of electrically stable MOS devices. By the end of 1960s electrical stability was achieved by using Phosphorous doped SiO_2 often known as P-Glass.

1970s saw the emergence of NMOS technology and NMOS based Dynamic Random Access Memory. IBM and Intel provided the technology and MOSTEK gave the Clock Design. This led to the successful marketing of the first VLSI 16K-bit DRAM. Scaling was applied to this DRAM structure.

A consortium of five top Japanese microelectronics companies: Hitachi, NEC, Fujitsu, Mitsubishi and Toshiba worked on VLSI Technology Project and created the complete infrastructure for 256K-bit DRAM . For this they developed Ultra-Clean Rooms in which they ushered in 1-micron VLSI era by the end of 1970s.

By 1980s NMOS circuits were replaced by CMOS due to ease of circuit design. Quasi constant voltage scaling was adopted. Voltage tolerant device structures like the lightly doped drain (LDD) transistor, silicide clad source drain, and hot electron defense was adopted. These technologies provided some of the keys to continue scaling feature sizes slower than voltage . 1980s saw the crossing of 1 micron barrier and formally we entered into sub-micron technology era.

In 1980s only Carver Meade and Lynn Conway wrote a book by the title “Introduction to VLSI System”. In this they simplified the design rules of CMOS circuit by introducing Linear Device Scaling Theory. University of California, Berkeley, introduced ‘CADENCE’- a suite of design tools of CMOS circuits. This connected the device physics to logic circuits, logic circuits to circuit design and verification tools to check the design rules. The entire ASIC world of semi-custom chips opened up based on this set of abstractions and made scaling applicable to all chips.

In 1990, Stanford developed process simulators and device simulators. They incorporated the scaling in Generation of Process Technology and a given Generation could be used for implementing a given system. Complex Systems would be implemented by advanced Generations. Thus emerged Technology Computer-Aided-Design today popularly known as TCAD.

In order to make TCAD commercially viable, manufacturing yield and defect analysis had to be included. Redundancy and repair techniques were included. Yield analysis tool (developed at Carnegie Mellon University), defect measurement tools, systematic yield process and ramp process kept ‘linear scaling’ commercially viable. This enabled scaling beyond 256K-bitDRAM.

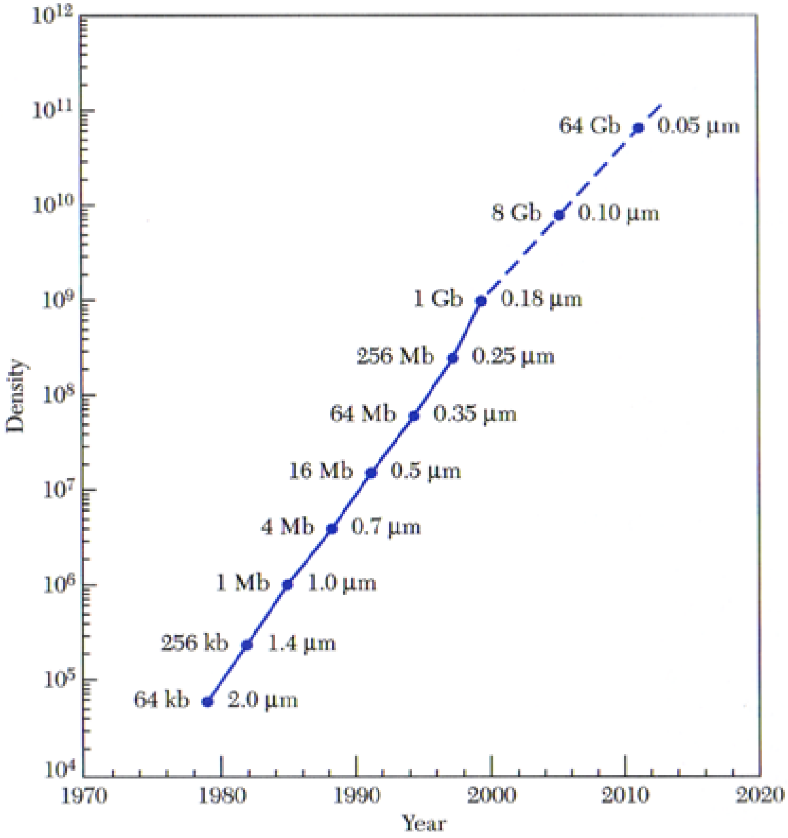


Figure 6.29. Scaling of Memory Chips.

As seen in Figure 6.29, from late 1970 the 2µm Generation of Technology(64kb per chip) improved to 0.18µm Technology (1Gb per chip) in 2000 for memory chips. In the same period Generation of Technology for logic chips improved from 10K devices to 100M device-count per chip as shown in Figure 6.30. As is evident from the two figures, Memory-Chip has been full one Generation ahead of Logic-Chips as far as technology is concerned.

Because of the edge of Japan over USA in early 80s in sub-micron VLSI Memory chips, USA under the aegies of Defence Advanced Research Program Agency (DARPA) started Microelectronics Manufacturing Science & Technology (MMST) Program. DARPA made a consortium with Texas Instruments, US Air-Force, Semi-conductor Equipment manufacturers and University Researchers. Its purpose was the following:

1. To develop advanced IC-manufacturing Technologies which would enable dramatic improvements in process control, cycle-time and overall flexibility;
2. Continue scaling in deep-submicron level in a cost effective manner.

Under MMST program, technical feasibility of 100% single-wafer processing, dynamic/object oriented Computer-Integrated-Manufacturing(CIM), real time/model based process control, in-situ sensors, 95% dry processing and integrated mini environment was well demonstrated. Till 1993 it was 60% single-wafer processing and 40% batch processing was in vogue. But after 1993, 100% single wafer process tools are available.

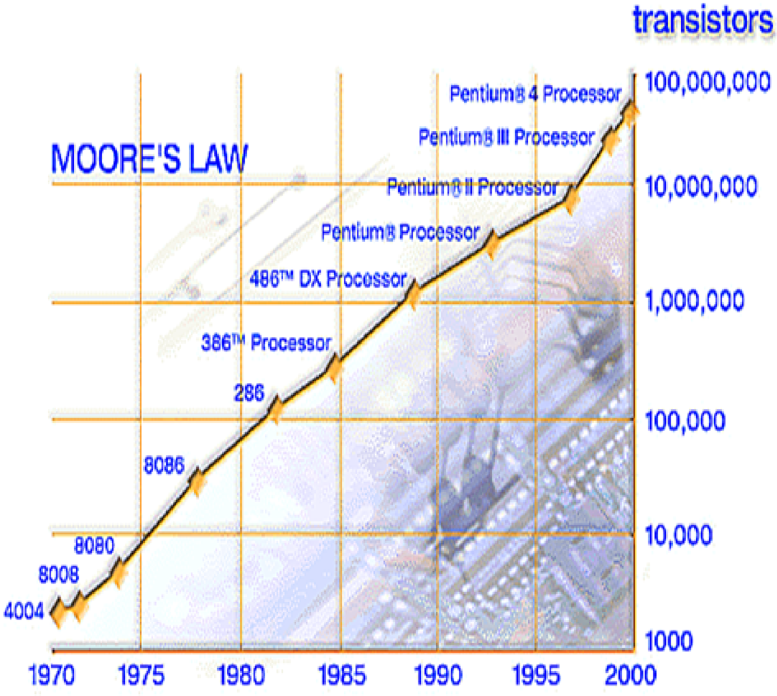


Figure 6.30 Scaling of Logic Chips.

MMST has developed processing chambers in which sample wafers were heated by Rapid Thermal Processing(RPT) Lamps under multi-zone, closed-loop wafer-temperature control. Stanford University, Applied Materials Inc., subsequently introduced RTP in their Centura cluster tools. MMST also created the first lithography cluster tool.

In 1992, Semiconductor Industry Association(SIA) invited 179 key Semiconductor Technologists at Irving, Texas, to chalk out the Road Map of scaling of MOS devices for the next 15 years as given in Table 6.8.

They identified 5 areas of critical challenge which could impede or stop the progress of scaling.

- Patterning material and processes for device structures below 0.25um
- Electrical interconnections, both on and off chip
- Electrical test, time cost and capability
- Design, modeling, simulation capability for all elements of IC technology and products
- Software capability, availability and quality for all aspects of IC technology and production.

Table 6.8_1992 SIA Overall Roadmap Technology Characteristics

	1992	1995	1998	2001	2004	2007
Feature Size (µm)	0.5	0.35	0.25	0.18	0.12	0.10
Gates/ chip	300k	500k	2M	5M	10M	20M
Bits/ Chip - DRAM - SRAM	16M/4M	64M/16M	256M/64M	1G/256M	4G/1G	16G/4G
Wafer	$4.00	$.390	$3.80	$3.70	$3.60	$3.50

processing cost ($/cm2)						
Chip Size (mm2)-logic processor-DRAM	250*132	400*206	600*228	800*500	1000*700	1250*1000
Wafer Diameter (mm)	200	200	200-400	200-400	200-400	200-400
Defect Density(Defects per cm2)	0.1	0.05	0.03	0.01	0.004	0.002
No. of interconnect levels - logic	3	4-5	5	5-6	6	6-7
Maximum Power (watts per die)-high performance-portable	103	154	304	404	40-1204	40-2004
Power supply	5-3.3	3.3-2.2	2.2-2.2	2.2-1.5	1.5-1.5	1.5-1.5

Voltage (V)-desktop-portable						
No. of I/O’s	500	750	1500	2000	3500	5000
Performance (MHz)-off chip-on chip	60/120	100/200	175/350	250/500	350/700	500/1000

Before 1980, there used to be Integrated Device Manufacturers(IDM) which were vertically integrated. They designed , fabricated and marketed their chips. Start-up companies, who could not afford a foundary, used IDM’s excess capacity to fabricate their chips. But after 1980 as Process Technology Scaling became more systematic, R&D and Design got bifurcated from the Foundries. A few major Foundries, situated in third world undeveloped or underdeveloped, catered to the needs of numerous Fab-less companies which concentrated their resources on the end market. Taiwan Semiconductor Manufacturing Corporation (TSMC) is one such foundries which enables a lower capital cost because of cheap labour. This allows the Fab-less companies to concentrate on their R&D. This makes a lot of economic sense.

The specific technology development challenges of a successful foundry company in the next decade include:________________________________

(1) aggressive scaling of transistors, interconnect, and design rules for both performance and density;__________________________________

(2) embedded technologies for SOC(System-on-Chip) solutions;____________

(3) cost effective and manufacturability process technology;______________

(4) a versatile technology portfolio;________________________

and (5) easy integration among customers, design service/IP providers and the foundries.___

In the following years as the Foundries build their core competencies by including leading edge process technologies, advanced and flexible manufacturing capabilities and customer oriented service systems, the dichotomy between the Foundries and the Fabless Companies will be complete.

As the scaling continues we need to answer the following challenges up to 90nm Generation Technology:

1. Sub-micron optical lithography(including Optical Proximity Corrections/Resolution

Enhancement);
2. Extending bulk CMOS by several performance boosters- stress/strain, ultra shallow junctions and ox nitrides;
3. Multi-level Cu interconnects including Chemical Mechanical Polishing(CMP).

At 45nm Generation Technology and beyond the following challenges have to be

answered:

1. Device/process variability: In the micron Generation the yield loss was dominated by random defects in the wafer. By the time volume production started by suitable quality control of wafers, systematic yield losses was eliminated. But in sub-micron and deep submicron Generation Technology, random defect limited yield losses are less than 50%. Major losses come from lay-out systematics and parametric systematics. For 90nm NMOS Generation Technology, systematic variation in line width is 40% of the overall chip variance. The systematic variation in the line width has a serious effect on the electrical performance which in turn effects the yield. This layout systematics and parametric systematics have to be addressed by the circuit designer. The histogram in Figure 6.31 illustrates the above point.

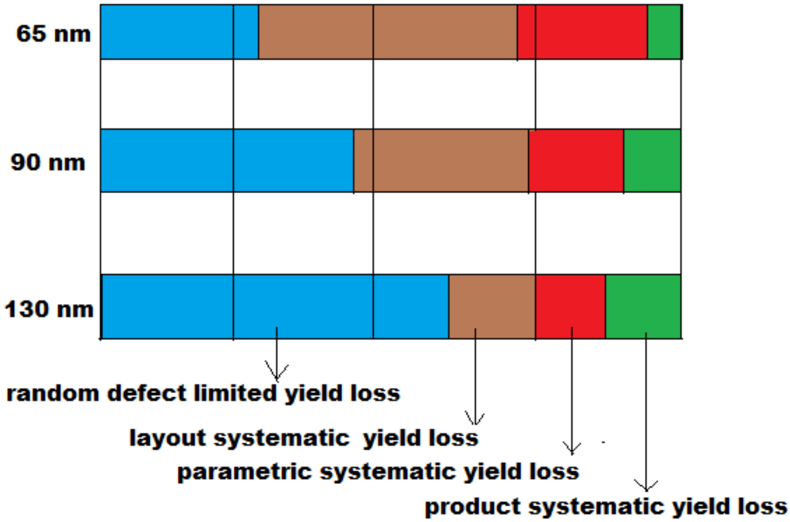


Figure 6.31

1. Despite of new device architecture such as FinFET, Ultra-Thin Body Transistors(UTB), Inverted T FET, the bulk CMOS will dominate. With scaling, performance parameters become sensitive to random dopant fluctuations. Circuit designers will have to cope with much higher variability in parameters as well as higher leakage currents.
2. With scaling due to Fermi-Level pinning (AppendixIV), higher threshold Voltage (V_{Thres}) resulted. Also High-k material have high elasticity hence result in higher phonon scattering or lattice scattering resulting in lower channel mobility. By screening the phonon effect, the deterioration in channel conductivity could be prevented. This required increasing the electron density in Poly-Si gate.

Hence we had to switch to Metal-High k combination. This prevented:

1. Fermi-Level pinning hence threshold voltage normalized;
2. This normalized channel conductivity which had drastically deteriorated due to dipole vibration effect;
3. Metal Gate Electrode provides better bonding between Metal-Dielectric. Hence stable operation.

In mid-2003, Intel's Hillsbaro, Ore, Development Fab developed HK-MG CMOS. Intel's 130nm technology was used. Using Hafnium-Based Oxide and Metal Gate electrode following characteristics were achieved:

1. Low Threshold = 1V;
2. Negligible leakage current through Oxide;
3. High 2-D channel mobility.
4. Advanced Process Control(APC) especially feed forward: Given the increasing complexity and small process windows, yield variability is a very significant problem. New approaches for yield relevant SOC(System on Chip) and APC are needed to take advantage of the increased process observables due to in-situ equipment sensor/FDC deployment.
5. Compact device models: Below 100 nm, compact device models must accommodate

microscopic (i.e., "non-bulk") physical effects with minimal impact on overall computational complexities. BSIM has filled this role for many technology generations as the workhorse, both for model characterization and node-to-node technology predictions. It continues to have the confidence of industry and seems likely to remain in service (with the possible exception of RF) down to about 45 nm.

Compact models at 65 nm need to be urgently improved on account of the following:

(a) scalability of sub-threshold currents and output resistance from short-to-long channel lengths, due largely to lateral doping non-uniformities

(b) dependence of noise on voltage and geometry; i.e., considering 1/f noise dependence on random noise trap occurrences

(c) capabilities for handling geometrical statistical fluctuations which affect noise, threshold voltage and drive current.

1. The above problems become more severe at 45 nm, along with the following additional priorities:

(1) gate current scaling and dependences on novel (e.g., multi-layer) gate stacks,

(2) carrier mobility in the channel due to layout-induced stress/strain,

(3) statistical variations stemming from random dopant placements,

(4) ballistic transport of carriers in intrinsic channel and,

(5) quantum mechanical effects due to confinement in thin films.

In nutshell, Scaling theory has been the organizing principle of the progress of the semiconductor industry throughout three decades. It has created a framework for continued improvement in density and cost performance and facilitated the desegregation of the entire industry around design and manufacturing. Few concepts in our time have had as much influence on the economy.

In this section we have done a rapid review of the challenges which have confronted the continuous microminiaturization of IC chips. In the next few sections we will study the specific technological developments and innovations made to meet these challenges.

SSPD_Chapter 6_Part 3_concluded_The list of Foundaries
This is the concluding part of SSPD_Chapter 6_Part 3.This gives the comprehensive list of the foundries and their technological capabilities.

SSPD_Chapter 6_Part 3_Concluded_The list of foundaries catering to different generations of technologies.

EE Times_News&Analysis_Chipmakers capabilities ranked_Peter Clarke_1/24/2011_11:58 AM EST. There will be three foundry sources of leading-edge semiconductor technology by the end of 2011: Taiwan Semiconductor Manufacturing Co. Ltd. (TSMC), GlobalFoundries Inc. and Samsung Electronics Co. Ltd., with possible addition of Intel, according to new research from IHS iSuppli.

Those three could be joined by an Intel entry into foundry supply. Intel Corp. may decide to offer foundry services to design companies and fabless semiconductor suppliers that are incorporating the Atom microprocessor into their designs. It could lead to significant revenue growth for Intel as well as more favorable asset utilization. The strategy of providing the Atom design to TSMC and its clients did not meet the expectations of either TSMC or Intel in 2010.

Several other foundries, including United Microelectronics Corp. and Semiconductor Manufacturing International Corp., will be capable of providing near-leading-edge foundry manufacturing in high volume. At the end of 2011 theere will be no Japanese suppliers manufacturing at the 32-nanometer and smaller geometries. Conceivably, Japanese equipment suppliers could leapfrog their competitors with the appropriate funding. If this were to happen, 2011 has to be the year that Japanese companies come out of their shells and become active in the foundry business. The Japanese semiconductor industry could suffer greatly if it chooses to remain in the sidelines, said iSuppli.

Table 6.8.1. Chipmakers capabilities ranked.

130nm	90nm	65nm	45/40 nm	32/28nm	22/20nm
Altis Semiconductor	Dongbu HiTek	SMIC	SMIC	UMC	TSMC
Dongbu Hi Tek	Grace Semiconductors	UMC	UMC	TSMC	Global foundaries
Grace Semiconductors	SMIC	TSMC	TSMC	Global foundaries	Intel
SMIC	UMC	Global	Global	ST	Samsung

		foundaries	foundaries	Microelectronics
UMC	TSMC	Infineon	Renesas(NEC)	Intel
TSMC	Global foundaries	Sony	IBM	Samsung
Global foundaries	Seiko Epson	Texas Instruments	Fujitsu	
Seiko Epson	Freescale	Renesas(NEC)	Toshiba	
Freescale	Infineon	IBM	ST Microelectronics	
Infineon	Sony	Fujitsu	Intel	
Sony	Texas Instruments	Toshiba	Samsung	
Texas Instruments	Renesas(NEC)	ST Microelectronics		
Renesas(NEC)	IBM	Intel		
IBM	Fujitsu	Samsung		
Fujitsu	Toshiba			
Toshiba	ST Microelectronics			
ST Microelectronics	Intel			
Intel	Samsung			
Samsung				

SSPD_Chapter 6_Part 4_The instability of MOS in 60s and its soluion

SSPD_Chapter 6_ Part 4 deals with the instability of MOS in 60s and how this problem was solved.

SSPD_Chapter 6_Part 4_The instability problem of MOS devices and its solution

In 1965 a very important paper was published "Investigation of Thermally Oxidized Silicon Surfaces Using Metal-Oxide-Semiconductor Structures" in *Solid State Electronics* in 1965. Due to electrical instability of Gate Oxide in MOS devices, commercial MOS ICs were not available in 60s. To overcome this instability, extensive studies were made of gate oxides using C-V measurements method.

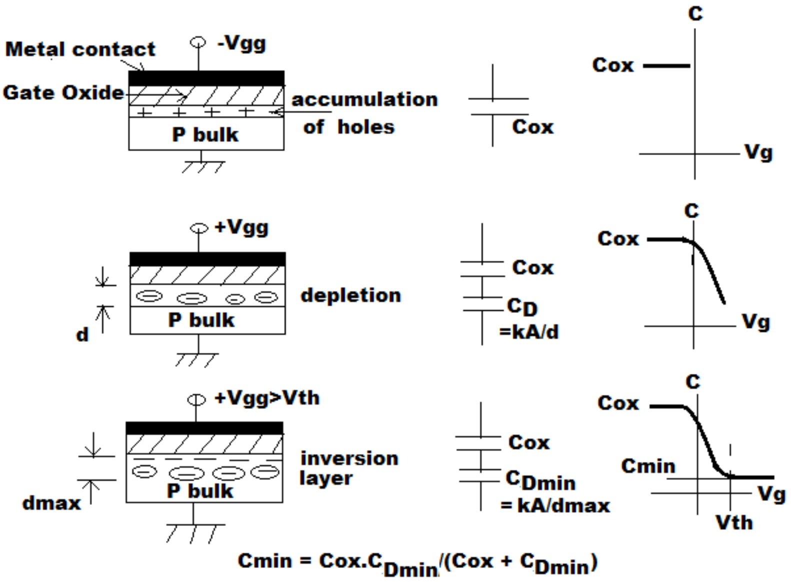


Figure 6.32 MOS C-V measurement method under steady state DC gate voltage bias.

MOS capacitance is Aluminum Gate contact deposited on dielectric SiO_2 thermally grown on P type Si-substrate. Here we assume that there are very little charges or defects in the oxide or at the interface of Si/SiO_2 or in the P type Si-Substrate.

There are four kinds of charges/defects present in Si/SiO_2 system which have been identified through C-V measurement curves namely:

1. Fixed Oxide Charge (Q_f) – this is about 10^9 to 10^{11}cm^{-2}. This lies in the gate oxide 2nm distant from the interface. It likely to be associated with the transition from Si to SiO_2 .

It is positive fixed charge and does not vary during the operation of the device.

2. Interface trapped Charge (Q_{it}) also known as surface states – this is possibly the incompletely oxidized Si atoms creating unsatisfied or dangling bond in the oxide layer but very near to the interface. They act as traps which can trap the mobile carrier of the channel. It will trap electrons in NMOS hence act as (-)ve charge and it will trap holes in case of PMOS acting as (+)ve charge. These act very much like bulk deep-level traps. The energy associated with the interface traps lie all along the forbidden gap but more so near the conduction band and the valence band. Oxidizing Si causes $Q_{it} = 10^9$ to $10^{11} cm^{-2} eV^{-1}$. This is of the same order as fixed oxide charge and it is experimentally established that both have the same origin hence higher fixed oxide charge implies higher interface trapped charges.
3. Mobile Oxide Charge (Q_m) – these are generally sodium (Na^+) and potassium (K^+) ions. These were the main culprits which were making MOS ICs technically unviable in 60s. These mobile charges were causing the instability in the threshold voltages. These were controlled by Phosphate doped SiO_2 (P-glass) and by replacing the tungsten filament for electron beam deposition of Aluminum Contact. Tungsten filament was the main

source of Alkali ion contamination. It was also experimentally established that the shift in V_{th} is inversely proportional to C_{ox} hence in subsequent years with scaling as C_{ox} increased the instability of the threshold voltage became less of a problem.

4. Oxide trapped Charges (Q_{ot}) – these are defects in the oxide layer which can trap electron or holes or both. These defects are caused due to broken Si and O bond due to ionizing radiation or due to process steps such as plasma etching or ion-implantation. These can be annealed and thereby minimized. But if they are not repaired they will cause trapped charges which will effect the threshold voltage of the MOS device. With scaling there is hot carrier effect and injection of hot carriers into oxide which may be trapped in these defects or the injected hot carriers may themselves cause defects leading to trapping of charges. These trapped charges will cause threshold instability.

Case 1. When gate voltage is (-) ve we get accumulation layer at the interface of Si/SiO_2.

The negative charge on the gate (Q_g) is exactly balanced by the accumulation of holes (Q_a) in the substrate just below the oxide layer. There is band bending in the Si-substrate due to the negative gate bias. This band bending results in valence band coming nearer to the Fermi-level as shown in Figure

6.33. This results in P-type semi-conductor becoming more P-type resulting in accumulation layer of holes just under the oxide layer. If at high frequency from (100kHz to 1MHz) the small signal capacitance is measured , the capacitance value is :

$$C_{ox} = (\varepsilon_0 \varepsilon_{oxide})A/(t_{oxide}) \qquad (6.6)$$

The measured value is given in upper part of Figure 6.32.

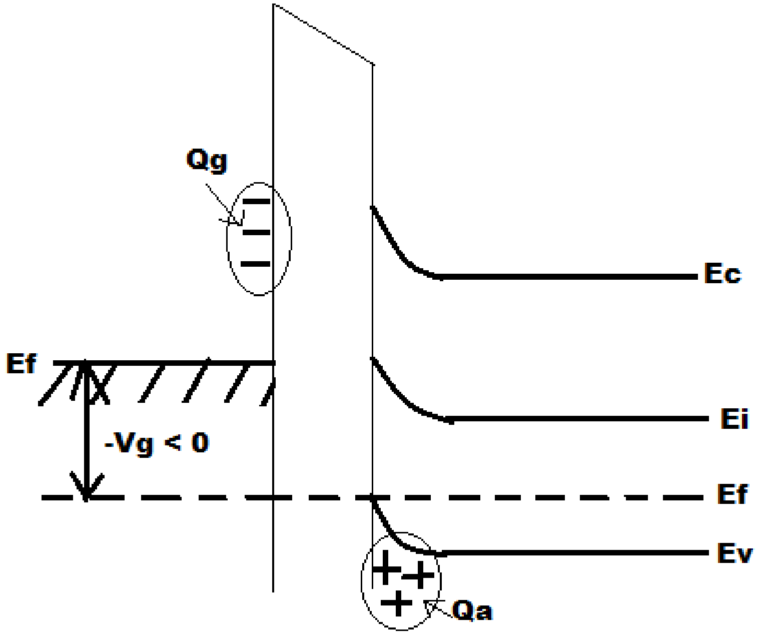


Figure 6.33 Creation of Accumulation layer below the oxide layer in the Si-substrate under (-)ve gate bias in an ideal MOS diode where

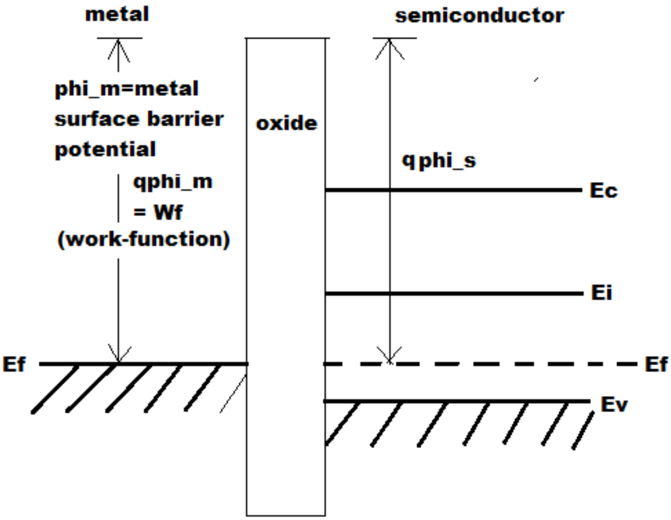
energy band is flat when gate bias is zero.

Case 2. When gate voltage is zero we get flat band condition since we have assumed ideal MOS diode.

Ideal MOS Diode has the following properties:

1. Energy Band is flat under no bias condition as shown in Figure 6.34;
2. Only charges are in the channel and in the gate and they are equal and opposite;
3. No carrier transport.

As seen in Figure 6.34, there is no band bending and if the small signal capacitance is measured at HF we get slightly less than C_{ox} as indicated in the middle diagram of Figure 6.32.


metal
semiconductor
phi_m=metal surface barrier potential
oxide
qphi_s
qphi_m = Wf (work-function)
Ec
Ei
Ef
Ef
Ev

In an ideal MOS diode there is no band bending. In real diode this is not the case.

Figure 6.34 Ideal MOS diode under zero bias condition.

Case 3. When gate voltage is (+) ve biased then the holes in the substrate nearest the interface are pushed away from the interface and depletion layer ‘d’micron wide is formed. The positive gate charge is balanced by the exposed (-) ve ionic immobile charge in the depletion layer. This (-) ve ionic immobile charge is due to negatively charged Boron atoms which have created the holes and in the process, by acquiring electrons, have become negatively charged.

The positive gate charge $Q_g = Q_d$ the depletion layer negative charge.

Now if small signal capacitance is measured at HF we measure the series combination of C_{ox} and C_D

where $C_D = [\varepsilon_0 \varepsilon_{Si} A/d]$ and d = depletion width______________________(6.7)

The measured $C = [C_{ox} \times C_D /(C_{ox} + C_D)]$__________________________(6.8)

The measured value of C is shown in the middle diagram of Figure 6.32

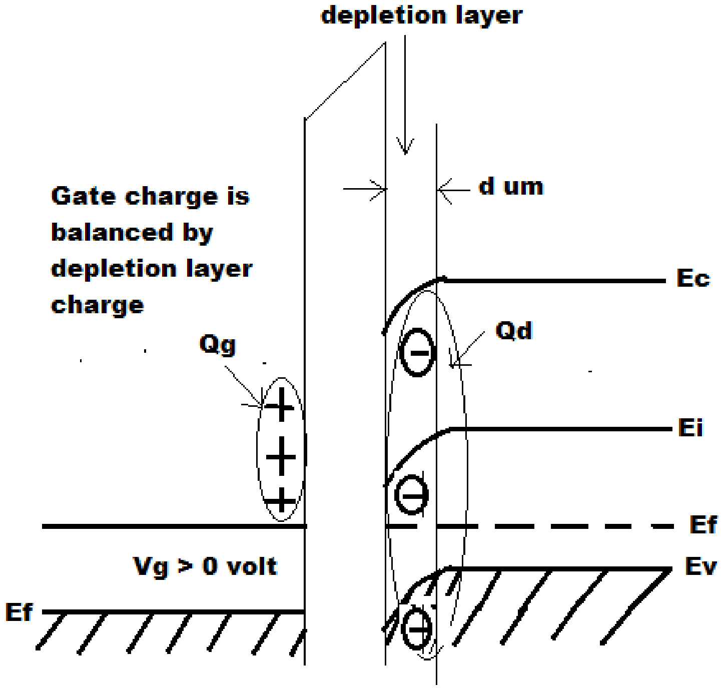


Figure 6.35 MOS Diode under positive gate bias.

Case 4. When Gate bias voltage exceeds threshold voltage.

The depletion layer will go on widening until threshold voltage is crossed. At the crossing of the Threshold Voltage , the layer of the substrate below the oxide layer is inverted as shown in Figure 6.36. P-type becomes N-type over a thin layer called inversion layer. A 2-D sheet of electron channel is formed just underneath the oxide later amounting to Q_i . Now the increase in gate bias voltage causes the electron density to become more copious. That is the channel becomes more conductive. But the depletion layer gets pinned at d_{max}.. In other words increase in gate charge Q_g is balanced by increase in Q_i in steady state condition or in slow variation condition.

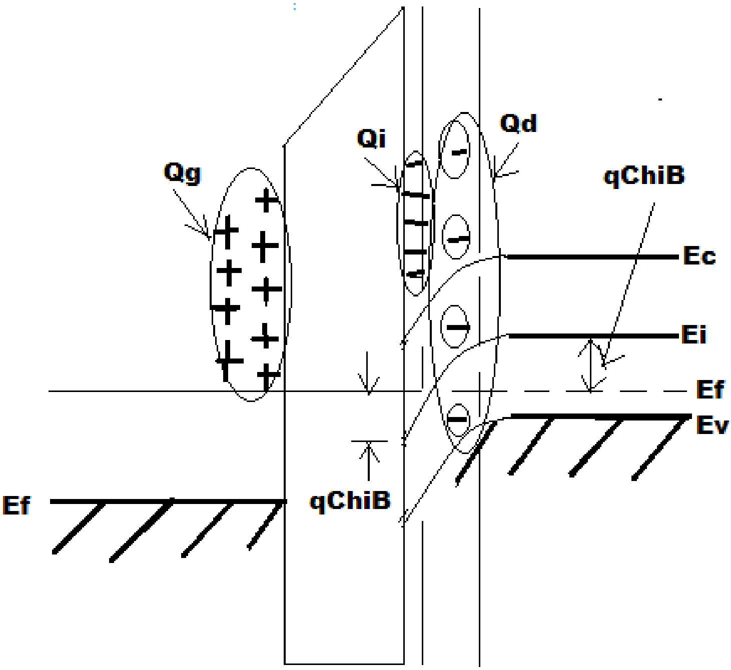


Figure 6.36. Ideal MOS diode with gate bias exceeding threshold voltage.

The n type channel has its electrons thermally generated hence its response is slow. Therefore when small signal capacitance is measured at HF then the rapid variation in gate charge can be matched only by the rapid variation in depletion width d_{max}. and hence ΔQ_g is balanced ΔQ_d .

Therefore the capacitance measured is:

C_{min} = series combination of C_{ox} and C_{Dmin} = [(C_{ox} × C_{Dmin}) / (C_{ox} + C_{Dmin})]_____(6.9)

Where C_{Dmin} = = [$\varepsilon_0 \varepsilon_{Si} A/d_{max}$]______________________________(6.10)

But under the same DC gate voltage if LF(Low Frequency) small signal capacitance is measured then ΔQ_g is balanced by ΔQ_i instead of ΔQ_d because under low speed variation condition or under steady state condition electrons of the inversion layer can easily be thermally generated or recombined. Hence LF capacitance measured will be C_{ox} as shown in Figure 6.37

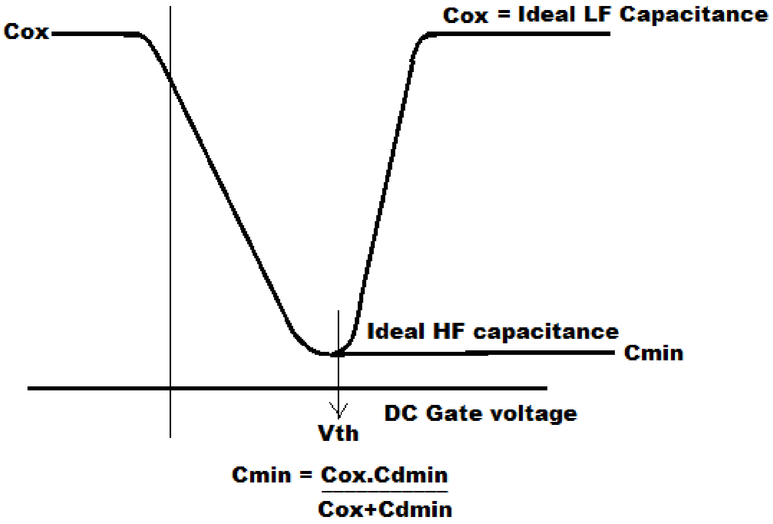


Figure 6.37 Ideal MOS capacitance curves at high and low frequencies.

But in reality , depending on the preparation conditions of the oxide, how it was formed and so on, will determine what electrical characteristics you have.In actual practice the position of the shape of C-V measurement in Figure 3.37 is not obtained.

The position of that capacitance drop along the voltage axis compared to a corresponding theoretical plot, is a measure of the number of charges in the oxide.

The more positive charges in the oxide, for instance, the point where the capacitance drops along the voltage axis will move to the left, to more negative values as shown in Figure 6.38. If you have, for instance, negative charges in the oxide, the curve will be shifted in the positive direction compared to the theoretical plot.

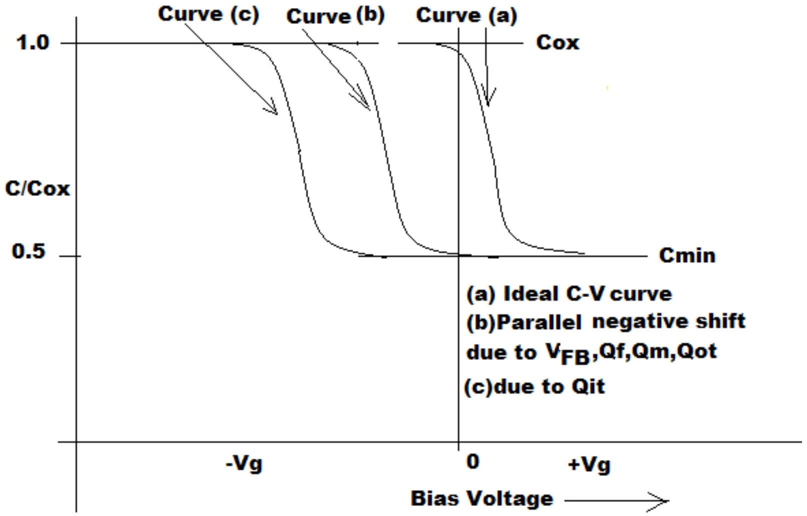


Figure 6.38 Departure or parallel shift from Ideal Curve (a) due to flat band voltage V FB , Fixed Oxide Charge (Q f), Mobile Oxide Charge (Q m), Interface trapped Charge (Q it), Oxide trapped Charges (Q ot)

We can measure the position of the C-V curve along the voltage axis before and after moving the sodium ions from the field plate interface to the silicon oxide interface using a bias-temperature stress test. You can move them back and forth and watch the C-V curve shift back and forth. This gives you a tool to measure the effect of all sorts of process conditions on charge formations. Stress conditions are normally ±50V/um of oxide, 300° C, and 2 minutes.

Table 6.9 Comparative study of density of various charges in two different orientations Si

crystals

	<100>	<111>
Qit	$10^{10}/cm^2/eV$	$10^{11}/cm^2/eV$
Qf	$10^{10}/cm^2$	$5\times10^{10}/cm^2$

All four types of charges have deleterious effect on device operation. During fabrication we choose process sequences which will minimize the charges. By high temperature inert anneals in Ar and N_2 towards the end of the process flow and a moderate 400°C anneal in H_2 or forming gas (N_2/H_2) at the end of the process accomplishes this.

The instabilities that had been noted over these years in thermal oxides and MOS transistors, were actually due to mobile ions, like sodium, lithium, potassium. Very interesting series of events were occurring, in that there were several laboratories working on the same problem. IBM, RCA, Bell Laboratories, Philco and a couple of others.

In '64 the PLANAR-II technology was introduced for the fabrication of stable MOS. The PLANAR-II was a term that was used to denote stable MOS.

In order to fabricate stable MOS we go through the following steps:

1.Using a clean metallization system, that is using electron-beam metallization, rather than filament metallization, which had been used previously.

2.To minimize the sodium incorporation into the process gettering steps were used. Here phosphorous is deposited on Silicon Dioxide to form phosphosilcate glass (PSG). This PSG is used to actually suck the impurity ions out of the oxide after they get in there. Impurity got in there after all of the attempts to keep them out.

To stabilize MOS device a vapor-deposited oxide is needed at the final stage of fabrication. This is needed for passivation and stable performance where the oxide is produced in the vapor phase and deposited over the device or circuit. We cannot do thermal oxidation as it is done after metallization. The process that was initially used was called TEOS, which was tetraethylorthosilicate.

A process which involved the breakdown of a organic oxygen-containing compound to form SiO2. This was done at about 750 degrees. The problem with this process is 750 degrees is way over the melting point of aluminum, and we desired to have something that would go down under 450 degrees, be deposited at less than 450 degrees. Silane, SiH4, can be reacted with oxygen, to produce SiO2. Again, a vapor-deposited process, but it would be deposited around 400 degrees. This helped achieve deposited oxide that would go over an integrated circuit, having metallized aluminum as the interconnect material.. Now it's quite a standard process. This standard process has added plasma, activated process with plasma, which produces a more even, uniform oxide.

Once the fabrication of stable MOS was achieved it became the brick for building all of the microprocessors.

SSPD_Chapter 6_Part 4_concluded_MOS Physics
SSPD_Chapter 6_Part 4_conclusion describes an ideal MOS diode, defines flat band voltage and gives the theoretical formulation of the threshold voltage.

SSPD_Chapter 6_Part 4_concluded_MOS and its Physics.

What is an IDEAL MOS DIODE ?

The Band-Diagram of an Ideal MOS Diode is shown in Figure (A).

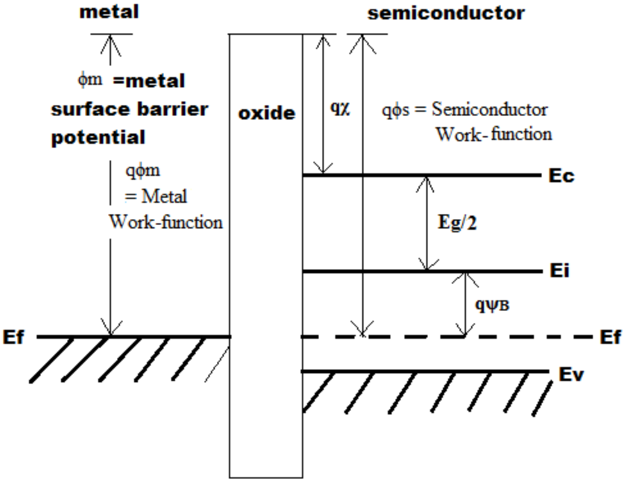

metal
semiconductor
ϕm =metal surface barrier potential
qϕm = Metal Work-function
oxide
qχ
qϕs = Semiconductor Work- function
Ec
Eg/2
Ei
qψB
Ef
Ef
Ev

Figure A. **In an ideal MOS diode there is no band bending. In real diode this is not the case.**

In an ideal MOS diode we have the following features:

1. The Flat Band Voltage = $V_{FB} = \phi m - \phi s$

where $q\phi m$ is the work-function of metal and $q\phi s$ is the work function of the semiconductor;

2. $V_{FB} = \phi m - \phi s = \phi m - [\chi + E_g/(2q) + \psi_B] = 0$ for P-type substrate where χ = electron affinity in semi-conductor, E_g = energy band-gap of Silicon and ψ_B = potemtial diiference between Fermi-level(Ef) and intrinsic level(Ei) in Si-Bulk;
3. Here the only charges are present at the gate and at the channel. These are equal and opposite.There is no charge in the oxide or at the interface of SiO_2/Si;
4. No carrier transport through the oxide layer.

In real MOS diode flat band voltage is not zero. There is always a difference between Metal Work-function and Semiconductor Work-function.

Consider an example of Metal (Al)-Oxide- P-typeSi as shown in Figure B:

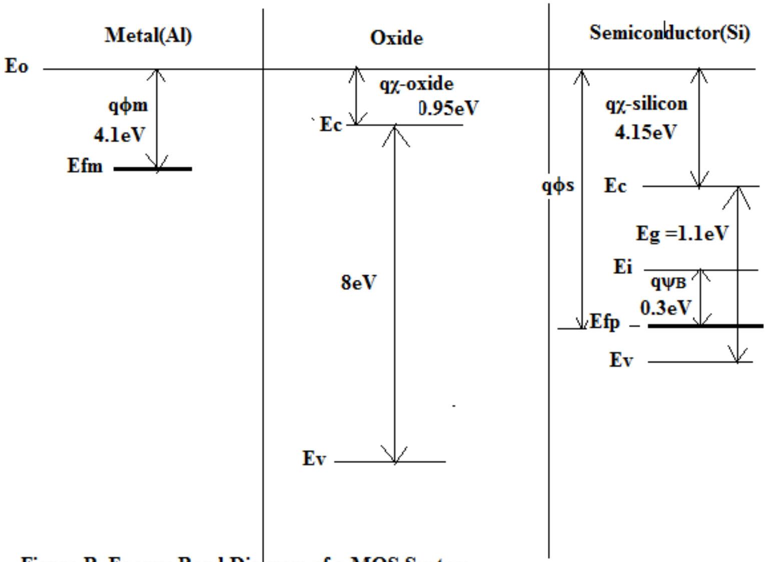


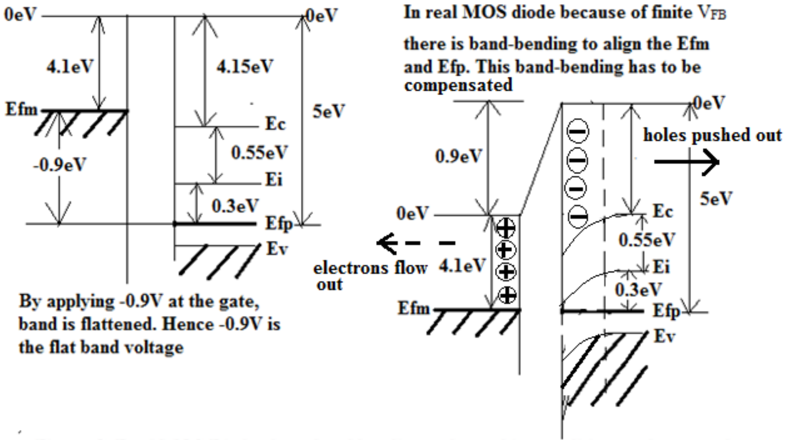
Figure B. Energy Band Diagram of a MOS System

The Flat Band Voltage is $= V_{FB} = \phi m - \phi s = \phi m - [\chi + E_g/(2q) + \psi_B]$

Therefore $qV_{FB} = q\phi m - q\phi s = q\phi m - [q\chi + E_g/(2) + q\psi_B] = 4.1 - [4.15+0.55+0.3]$

Therefore $qV_{FB} = 4.1 - 5.0 = -0.9\text{eV}$;

Because of this Flat-Band Voltage there is band bending even at zero gate voltage when gate and bulk are shorted. there is a depletion layer of negatively charged immobile Born ions just under the interface of SiO_2/Si as shown in Figure C. Because of band-bending near the interface Si-bulk is more intrinsic hence holes are pushed out leaving behind the space-charge region. Negative electrons leave Gate Plate and recombine with the holes pushed out from the bulk. Therefore we get a dipole layer centered on the oxide layer as shown in Figure C. Negative voltage -0.9 eV when applied to the Gate of MOS system compensates for this band bending and flattens the Band Diagram as shown on the left side of Figure C. On the right side of Figure C we see that below the interface in the Si-bulk we have near-intrinsic semiconductor. In effect a depletion layer exists in the bulk under the interface. If the band bending is $2\ q\psi_B$ then the depletion layer reaches its maximum width d_{max} and when band bending exceeds $2\ q\psi_B$ then an inversion layer is formed just under the interface in the bulk.



By applying -0.9V at the gate, band is flattened. Hence -0.9V is the flat band voltage

In real MOS diode because of finite V_{FB} there is band-bending to align the Efm and Efp. This band-bending has to be compensated

Figure C. Real MOS Diodes have band bending under no bias condition as shown on the right. To compensate for this band-bending , flat band voltge has to be applied as shown on the left. This flat band voltage is present in threshold voltage theoretical formulation.

Theoretical Formulation of Threshold Voltage:

We will see in Chapter 6, Part 6, that the theoretical formula for thrshod voltage is given by:

$$V_{Th} = V_{FB} + 2\psi_B + \frac{\sqrt{2\epsilon_S q N_A (2\psi_B)}}{C_{OX}}$$

6.A

Let us look at the ideal MOS diode in Figure A.

We have to apply sufficient gate voltage so the band bending at the interface is $2q\psi_B$ as shown in Figure D. This band bending is known as ‘Surface Potential Parameter’ = $q\psi_S$ = $2q\psi_B$. At this Gate to Source Voltage = V_{GS} we have applied sufficient voltage to just cross the Threshold Voltage ‘V_{Th}’ and turn on the n-channel. The 2-D layer just below the Si-SiO_2 has experienced INVERSION. P-type semiconductor has been inverted to N-type semiconductor. The charges induced, the electric field created and the potential drop caused from Gate-to-Bulk is shown in Figure E.

.

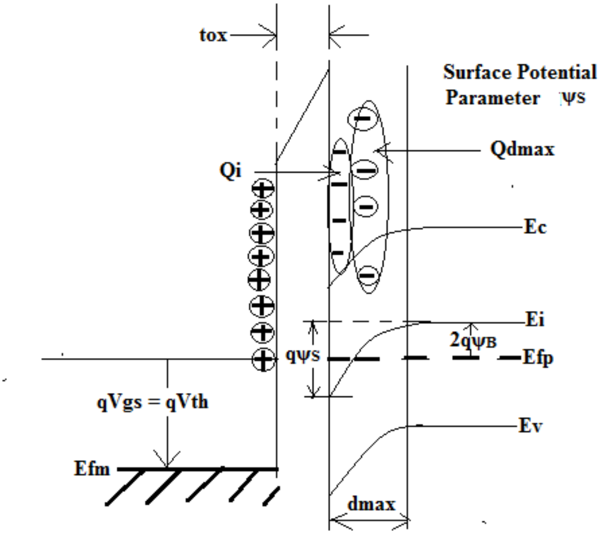


Figure D. Energy Band diagram of MOS Diode when Gate Voltage (Vgs) is equal to Threshold Voltage (Vth) and n-channel has just been turned on.

As seen in Figure E, the charge on the Gate (Q_G) is exactly balanced by maximum depletion layer space-charge (Q_{dmax}) + inversion layer charge (Q_i).

That is $Q_G = (Q_{dmax} + Q_i)$________________________
A.1

Because of electric flux lines originating on Q_G and terminating on ($Q_{dmax} + Q_i$) there is an uniform field in oxide layer E_0 and linearly graded E_1 just underneath the interface distributed across the depletion layer dmax wide.

$$E_0 = \frac{Q_G}{W \times L \times \epsilon_{ox}}$$

A.2

Where W = width of the channel and L= length of the channel.

$$E_1 = \frac{Q_{dmax}}{W \times L \times \varepsilon_{Si}}$$

A.3

Qdmax is the space charge of the depletion layer ‘dmax’ wide.

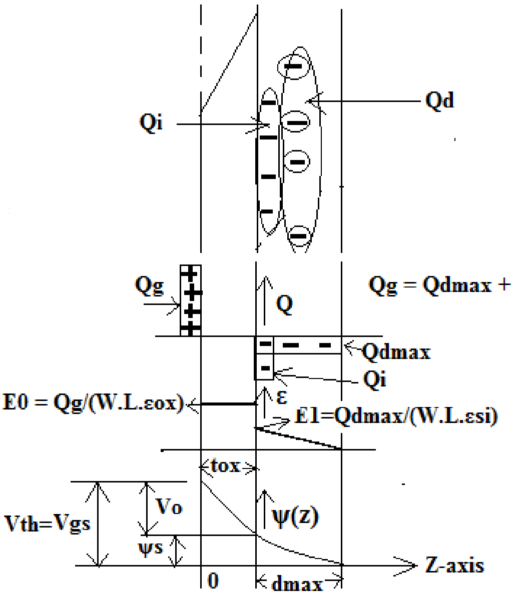


Figure E. The charge profile, Electric Field Distribution and the Potential Drop from the Gate to bulk along the longtitudinal axis-z is shown.

The depletion layer of PN junction diode :

$$W = d_p + d_n = \sqrt{\left[\frac{2\varepsilon_{Si}}{q}\left(\frac{1}{N_A} + \frac{1}{N_D}\right) \emptyset_{BO}\right]}$$

A.4

In MOS diode the depletion layer extnds in P-type bulk only and surface potential barrier (ψ_S) is equivalent to junction built-in potential barrier under no bias (ϕ_{BO}).

Assuming donor doping to be infinite and $\phi_{BO} = \psi_S$ equation A.4 reduces to:

$$d_p = \sqrt{[\frac{2\epsilon_{Si}}{q}\left(\frac{1}{N_A}\right)\psi_S]}$$

A.5

When $V_{GS} = V_{Th}$ then $d_p = dmax$ and

$$dmax = \sqrt{[\frac{2\epsilon_{Si}}{q}\left(\frac{1}{N_A}\right)2\psi_B]}$$

A.6

The space charge in depletion layer is:

$$Q_{dmax} = qN_A dmax$$

A.7

Substituting A.6 in A.7 we get:

$$Q_{dmax} = \sqrt{2qN_A\epsilon_{Si}(2\psi_B)}$$

A.8

Referring to Figure E, in an ideal MOS Diode just at the point of turning ON:

$V_{Th} = V_0 + 2\psi_B$;

Assuming Q_G is equal to the space charge (Q_{dmax}) in dmax depletion layer with inversion layer weakly induced we have :

$$V_{Th} = \frac{Q_{dmax}}{C_{ox}} + 2\psi_B$$

A.9

Substituting A.8 in A.9 we get:

$$V_{Th} = \frac{\sqrt{2q\,N_A\,\epsilon_{Si}(2\psi_B)}}{C_{ox}} + 2\psi_B$$

A.10

We have already seen that Flat Band Voltage, V_{FB}, has to be applied to flatten the band. This means a part of the job of the Gate Voltage is being done by V_{FB} . Hence if this has to be accounted then Equation A.10 becomes A.11.

$$V_{Th} = \frac{\sqrt{2q\,N_A\,\epsilon_{Si}(2\psi_B)}}{C_{ox}} + \; 2\psi_B + V_{FB}$$

A.11

If source to bulk voltage (V_{SB}) is not zero then this also has to be accounted. This is called Body Effect. Now Equation A.11 becomes:

$$V_{Th} = \frac{\sqrt{2q\,N_A\,\epsilon_{Si}(2\psi_B+V_{SB})}}{C_{ox}} + \; 2\psi_B + V_{FB}$$

A.12

If the charges in Oxide layer and at Si-SiO_2 have to be accounted then

Q_{SS} = Q_{it} (interface trapped charges) + Q_{ot}(oxide trapped charges) + Q_m(mobile chares in the oxide) + Q_f (fixed oxide charges) and Equation A.12 becomes:

$$V_{Th} = \frac{\sqrt{2q\,N_A\,\epsilon_{Si}(2\psi_B+V_{SB})}-Q_{SS}}{C_{ox}} + \; 2\psi_B + V_{FB}$$

A.13

This completes the theoretical formulation of the Threshold Voltage MOS device.

SSPD_Chapter 6_Part 5_Crossing of the Rubicon_Emergence of MOS Technolgy as the dominant IC Technology.

In SSPD_Chapter 6_Part 4 we saw the chemistry of MOS devices and the technology of achieving stable MOS devices. In Part 5 we will see what application specifics we must concentrate in order to make MOS technology commercially viable.

SSPD_Chapter 6_Part 5_ Crossing of the Rubicon_Emergence of MOS Technolgy as the dominant IC Technology.

("MOS Technology,1963-1974: A Dozen Crucial Years" by Ross Knox Bassett, The Electrochemical Society *Interface*, Fall 2007,pp 46-50)

An ancient Roman Law forbade any Roman General from crossing the river Rubicon and enter Italy proper with a standing army. In January 49BC an edict was issued by the Senate instructing the Governor of Gaul , Julius Caeser, to resign from his post and disband his army. Instead Julius Caeser defied the edict and with his standing army he crossed the Rubicon river and defeated Pompey, the Roman Emperor. This was a pivotal event and marked the emergence of Roman Empire and genesis of modern European Culture.

During the period 1963-66, the concerted efforts of Bruce Deal, Andrew Grove and Ed Snow at Fairchild and systematic works carried out by NEC, IBM and Phillips identified sodium contaminations as the source of instability and resolved the yield and reliability issues. But due to the slow speed of MOS devices which is inherent in 10μm long channel length, in 1965 IBM's Corporate Technical Committee issued a directive to IBM Research Group asking them to halt all works on MOS Technology. Instead IBM Research Group redoubled its efforts in application areas where MOS could be suitably applied. In 1966, an important breakthrough was made in the fabrication of memory cells at IBM. Till then six BJT was required to build 1-bit memory cell. But Robert Dennard showed that using MOS , one-transistor memory cell could be built. This was a great economy of real estate on Silicon Surface and large – size memory could be accommodated on a chip. The slow speed of MOS did not come in the way of commercializing a MOS memory chip. So MOS memories became a hot topic of Research in IBM. By 1970 MOS emerged as the dominant IC Technology. This crucial work done in 1963-66 , already described in the previous chapter, and the subsequent defiance of IBM Research Group to persist in MOS research is likened to "Crossing of the Rubicon". By mid-1970s MOS technology was established as a commercially successful and sustainable technology.

1964_General Microelectronics introduced first commercial MOS IC calculator. It had 23 custom IC

Chips.

1968_MOS Memories gave three-fold advantage in terms of packing density with respect to bipolar memories and also because of saturation in the performance of magnetic memories, IBM decided that their computer systems would use MOS memories after initial phase-in period using bipolar technology.

1968_Robert Noyce and Gordon Moore left Fairchild and established INTEL where they concentrated exclusively on MOS Technology for commercial production of memories and micro-processors. Subsequently this group was joined by Andy Grove.

1969_Rockwell Microelectronics was able to include all the functions in 4 MOS IC chips fabricated for SHARP's first portable calculator micro-compet QT-8D .

1971_Mostek and Texas Instrument introduced single chip calculator.

1972_IBM systems were introduced with 1024 bit MOS memory chips. At almost the same time MOS technology found an organizational home in IBM, Burlington, Vermont.

At Intel, it was decided to adopt self-aligning Silicon gate MOS Technology developed at BELL Labs. The self-aligning feature had a further packing density advantage. This technology helped produce stable devices in mass scale. Intel after an initial failure in marketing 256 bit MOS memory chip, successfully marketed 1-kbit MOS memory chip code number 1103 to Honeywell. The 1103 chip became a standard memory chip for non-IBM computer systems.

Besides MOS memory chips, Intel also successfully introduced MOS microprocessor chips and Erasable, Programmable ROM (EPROM).

1974_Intel introduced 4kbit MOS memory chips and second generation MOS microprocessor chip namely 8080.

By 1974 in a land mark paper by Robert Dennard et al, "**Device Scaling** Limits of Si MOSFETs **…** with Very Small Dimensions," IEEE Journal Solid State Circuits Vol 9, 1974, pp 256, the scaling principles were outlined. It was shown that scaling of dimensions by a factor scaled the time delay by the same factor and scaled down the power consumption as square of that factor.

Scaling became the treadmill for advancing the level of integration from micron to sub-micron level to deep sub-micron level and now we are advancing level of integration to ultra deep sub-micron level. Probably in next 10 years this treadmill of advancing the miniaturization will stop at nano level. Beyond this what alternative technology will become the treadmill for advancing miniaturization remains an unanswered question.This alternative technology could be graphene, spintronics or it could be genetics.Future will settle this question.

MOS Technology could reach its commanding heights first because of research in physics and chemistry MOS structures; second because product design and development to create ICs that had some advantages over bipolar ICs; and third because of organizational change to create environments where MOS technology could thrive. As we saw that the establishment of Intel went a long way in establishing the economic viability and technical feasibility of MOS Technology.

SSPD_Chapter 6_Part 6_MOS ICs Fabrication Technology_to be continued

SSPD_Chapter 6_ Part 6 introduces us to CMOS process flow upto to the fabrication of P-tub and N-tub.

SSPD_Chapter 6_Part 6_ MOS ICs Fabrication Technology

Because of the inherent advantage of MOS ICs, today MOS Technology dominates the IC market as compared to Bipolar Technology. Table 6.10 gives a comparative study of the two technologies.

Table 6.10 Comparative study of Bipolar and MOS Technologies.

	Bipolar Technology	MOS Technology
Carrier participation in transistor action	Both majority and minority carriers partake in transistor action bipolar	Only majority carriers partake in transistor action-unipolar
speed	Traditionally much faster	Slower devices

Power dissipation	Higher dissipation	CMOS is nano watt logic under standby condition
Packing density	One generation behind because of isolation diffusion	Much higher packing density
Design flexibility		Higher design flexibility
Noise Figure	Higher noise figure	Quieter devices.
Dynamic range	Input dynamic range is 5mV because it is exponential device.	Input dynamic range is 400mV because it is a quadratic device.

Semiconductor Industry Association(SIA) brought out the National Technology Roadmap for Semiconductors(NTRS) in 1994, updated in 1997 , again in 1999 and again in 2001. Based on that data we give the following Table 6.11.

Table 6.11. Future Projections for Silicon Technology taken from SIA NTRS (1997).

Year of the first DRAM shipment	1997	1999	2003	2006	2009	2012
Minimum feature size(nm)	250	180	130	100	70	50
Memory size(bits/chip)	256M	1G	4G	16G	64G	256G
DRAM Chip size(mm2)	280	400	560	790	1120	1580
µP transistors/chip	11M	21M	76M	200M	520M	1.4B
Maximum wiring levels	6	6-7	7	7-8	8-9	9
Minimum mask count	22	22-24	24	24-26	26-28	28
Minimum supply voltage	1.8-2.5	1.5-1.8	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6

Table 6.12 Updating of 1997 SIA NTRS Roadmap in 1999 and 2001.

	2004	2005	2007	2008	2010	2011	2013	2014	2016
1999_NTRS		100nm		70nm		50nm		35nm	
2001_ITRS	90nm		65nm		45nm		32nm		22nm

By a variety of post-lithographic techniques, the physical gate length is shorter than the printed size in photo-resist. It is projected to be 9 nm by 2016. "Lithography half-pitch and transistor gate length scaling trends continue to accelerate. This means that semiconductor chips will continue to get smaller, faster and ultimately less expensive at an even faster rate in the future," noted Paolo Gargini, Chairman of the International Roadmap Committee and Fellow, Intel Corporation "When the 2001 Roadmap looks 15 years into the future, the physical gate length is projected to be a mere 9 nanometers. We are beginning to consider technologies beyond planar or even post-CMOS devices."

In 1970, NMOS IC Technology was introduced but it was replaced by CMOS IC Technology by 1980. So our discussion will concentrate on CMOS Process Flow.

Section 6.6.1. NMOS IC in 1970s.

For academic reasons, we introduce the cross-section view of NMOS IC in Figure 6.39. This was prevalent in 1970s. This includes both enhancement type NMOS and depletion type NMOS.

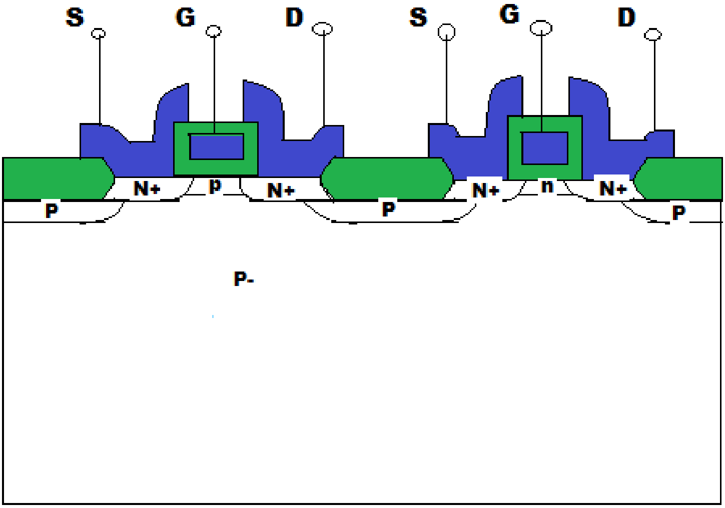


Figure 6.39 1970s NMOS IC. Left MOS is enhancement type NMOS and right is depletion type NMOS

In Figure 6.39, blue colour represents METAL CONTACT and green colour represents OXIDE. Thin oxide is gate oxide and thick oxide is field oxide.

Section 6.6.2. CMOS process flow(“Silicon VLSI Technology-Fundamentals, Practice and Modeling” by James D. Plummer, Michael D.Deal and Peter B. Griffin, Pearson Indian Edition 2009).

The ease of circuit design and the fact that the standby power is of nanowatt level compelled VLSI Circuit designers to adopt CMOS as the building block from 1980 onward. The trend continues to be the same and it will remain so for another decade. Hence while describing VLSI process flow we will concentrate on CMOS process flow. CMOS can be used an active device or it can be used as a capacitance or resistance in VLSI chips.

Section 6.6.2.1. Choice of a substrate.

Just as we chose a P-type substrate of doping 10^{15}Boron atoms per cc of resistivity 10Ω-cm of crystal orientation <111> for MC1530 fabrication in the same way we will choose P-type substrate of doping 10^{15}Boron atoms per cc of resistivity 15 Ω-cm but of crystal orientation <100>. In crystal orientation <100> as we saw in Chapter 6, Part4, the fixed oxide charges, mobile oxide charges, interface trapped charges and oxide trapped charges are one order of magnitude less than those in crystal orientation <111> . Also threshold voltage is minimized in crystal orientation <100>. Since the fixed and mobile charges in the oxide layer or at the interface have to be minimized for electrical stability of the threshold voltage hence we prefer crystal orientation of <100>.

“Intrinsic gettering” is done to minimize the sensitivity of the wafers to trace contaminants which may be introduced in subsequent processing. Idea of gettering is to confine the alkali ions and metal ions to the back portion of the substrate where they are well removed from the active regions of the substrate.

While making CMOS VLSI IC there are three options as shown in Figure 6.40:

Option 1- start with P-substrate with N tub;

Option 2-start with N-substrate with P tub;

Option 3- start with P-substrate with twin tubs.

If we adopt option 1 then substrate will contain NMOS and tub will contain PMOS. If we adopt option 2 then substrate will contain PMOS and tub will contain NMOS. If we choose option 3 then the two tubs will be N type and P type each of 10^{16} to 10^{17}dopent per cc.

Generally twin tub approached is adopted because that gives a greater flexibility in design and better control over the dopent profile of the active devices since tubs are fabricated by ion-implantation which is far better in terms of the control and reproducibility of the dopent profile and the pn shallow junction. Here we may briefly discuss “ion implantation” and its superiority over “ thermal diffusion” in terms of reproducibility and control over the dopent profile.

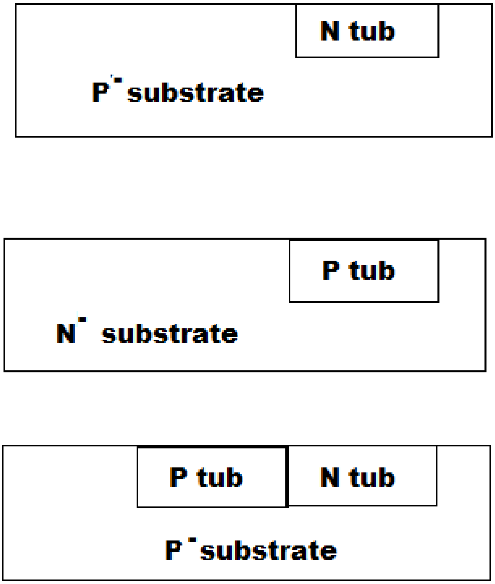


Figure 6.40 Three options for building CMOS VLSI Chip.

Section 6.6.2.2 Ion Implantation

With generation to generation of technology as lateral scaling has taken place so has the vertical scaling has taken place. This has led to the need of very precise and reproducible shallow junction depths. This need cannot be fulfilled by conventional thermal diffusion. So ion-implantation method was adopted and today in VLSI/ULSI era this is the dominant doping technique for both MOS as well as Bipolar technology. An ion-implantation chamber is shown in Figure 6.41.

In this technology, mass is very critical because mass decides the average Gaussian profile of the dopent. For a given dopent we must choose a given isotope and the most abundant isotope of the dopent. Hence mass analyzer is very important. Also all the dopent must be ionic and not neutral. This is because ionic dopent can be counted and quantitative measure of the dose can be maintained. This will help in reproducibility of the dopent profile. So it is very important that neutral atoms are removed from the ion beam. Therefore ion beam undergoes electrostatic deviation from the straight line path before final implantation. This acts as a trap for the neutral dopent atoms which continue on a straight line path.

The implant dose is measured by putting the wafer deep inside a Faraday Cup. This cup collects the ion current and integrates it over time.

$$Q = \frac{1}{A} \int \frac{I}{q}\, dt$$

6.11

Where I = collected beam current, A = implant area, t = the integration time and q is the charge on the ion.

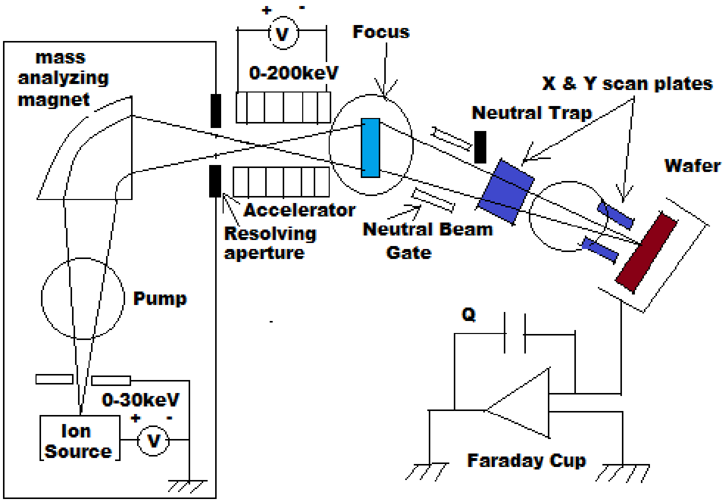


Figure 6.41 Schematic of an ion-implanter.

Conventional ion implanter has much in common with linear accelerators. The ion source can be a solid elemental dopent source or it could be in compound gaseous form such as Arsenie, Bromine or Phosphine diluted by 15% Hydrogen. These are ionized either by electrons due to thermoionic emission or due to plasma gas discharge.

The ions are extracted by a grid at 30kV. Next it is passed through a mass analyzer. In mass analyzer it undergoes magneto-static bending as defined by the following formulation:

$$\frac{m \times v^2}{R} = q.v \times B$$

Where m = mass of the ion; v = velocity of the ion, B = magnetic field density = α.I where I is the current in the magnetic coil.

Energy imparted to the ion while being accelerated through extraction Voltage on the grid: $E = q.V_{extraction}$ 6.13

Therefore the final velocity of the ion after emerging from the grid is:

$(1/2)mv_2 = q.\ V_{extraction}$

Or $v = \sqrt{(2q\ V_{extraction}/m)}$ 6.14

From equations 6.12 and 6.14 we get:

$$\sqrt{m} = \frac{q}{\sqrt{2E}}(\alpha R)I$$

6.15

From Equation 6.15 we see that ions coming from the resolving aperture will be of mass m decided by the current I flowing through the magnetic coil of the mass analyzer. Thus by tuning the current ‘I’ we can select the ions which will exit the resolving aperture and constitute the beam current.

As seen in Figure 6.41, after the resolving aperture the ions are further accelerated through a voltage of 0 to 200kV. This accelerated beam of ions are smashed into the wafer. Each impacting ion dislodges thousands of Si atoms and damages the Silicon surface. The total number of ions implanted in the Silicon target near the surface are called the Dose of Implant. This dose is given by Equation 6.11. Hence total dose is the product of ion current and the time of implant.

The standard dose of implant is 10^{12}/cm^{2} to 10^{16}/cm^{2}. Lower or higher dose is possible. By controlling the dose of implant the dopent profile can precisely be tailored according to our need.

The depth of implanted ion is random but for a large dose of implanted ions, the doping profile is a Gaussian Distribution as shown in Figure 6.42. The heavy ions like Antimony has a very shallow penetration while light ions like Boron have a deep penetration.

The average range of penetration is R_P = this is the range below the surface of the wafer where maximum number of ions are implanted and this ion peak concentration is given by C_P .

The concentration at any range ‘x’ is given by:

$$C(x) = C_p \exp\left(-\frac{(x-R_P)^2}{2\Delta R_P^2}\right)$$

6.16

The dose of implant is $Q = \sqrt{(2\pi)}\Delta R_P C_P$ 6.17

Here ΔR_P = the standard deviation or straggle about the average range R_P .

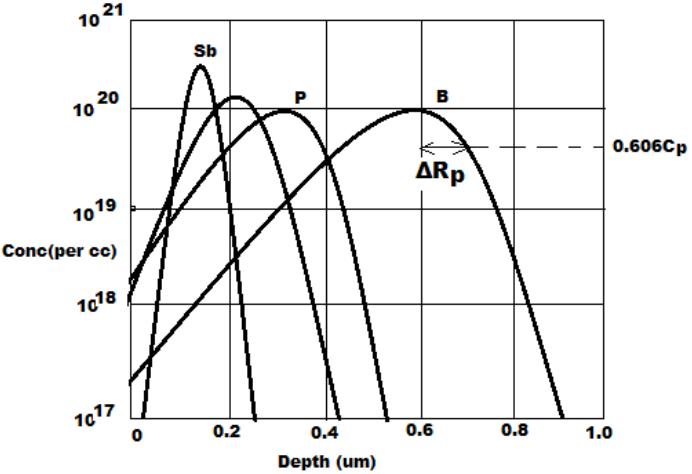


Figure 6.42 Distribution of ions implanted into crystalline silicon at 200keV kinetic energy. The light ions travel further and have broader distribution than the heavy ions.

After the implantation the damage inflicted to the silicon surface is annealed at high temperature. By annealing the damage gets repaired. The dislodged Si atoms diffuse to a vacant lattice site and fills up the defect. The annealing cycle is carried out at 1000°C for 10 seconds or at 800°C for 30 minutes.

During the annealing process the ion implant diffuses out but maintains the Gaussian distribution as shown in Figure 6.43

During ion-implantation, a thick layer of photo-resist (1µm) is an effective mask for shielding the areas where implantation is not desired. This is because ion-implantation is carried out at room temperature.

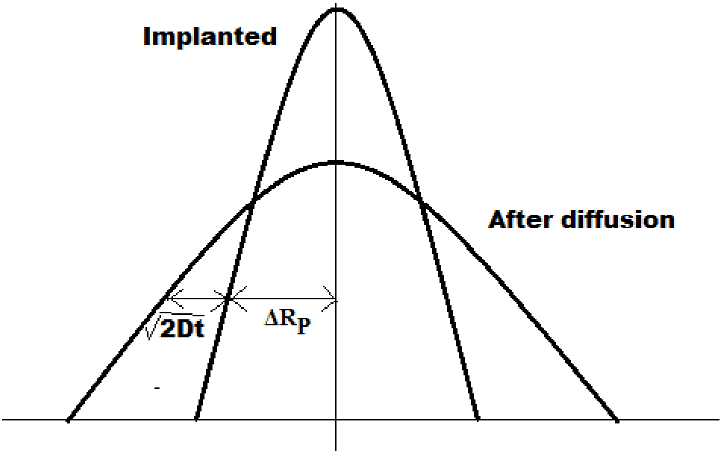


Figure 6.43 Evolution of Gaussian profile after annealing. The Gaussian preserves its shape as it diffuses in an infinite medium.

Section 6.6.2.3 Active Region Formation.

In bipolar technology we did isolation diffusion for isolating the active devices. Here we grow field oxide to isolate the MOS devices. The field oxide is grown by a technique called LOCOS described in Section 6.4.3 and in case of ULSI circuits we use Poly buffered LOCOS as described in Section 6.4.5.

As already discussed in Section 6.4.3 LOCOS suffers from bird’s beak problem. To circumvent this problem we adopt Shallow Trench Isolation. Instead of growing Field Oxide , in place of Field Oxide we etch out a shallow 500nm deep trench in Silicon, then we grow a thin layer of Oxide and fill it up with Silicon Dioxide. The thin layer of oxide is grown on the bottom and side of the trench by thermal oxide. This step reduces the charges at the Si/SiO_2 interface. This gives a vertical edge to the trench and removes the problem of bird’s beak which reduces the active area which cannot be afforded at VLSI/ULSI level. Next by CVD oxide is filled up in the trench and by Chemical Mechanical Polishing(CMP) planarization of the surface is done. This CMP process stops at Silicon Nitride.

The etchant of Silicon is Nitric Acid plus Hydrofluoric acid and the reaction is as follows:

$$Si + HNO_3 + 6HF \equiv H_2SiF_6 + H_2O + H_2$$

____________________6.18

Here the reaction takes place in two steps: Nitric Acid partially decomposes into NO_2 and oxidizes Si and then HF acid etches SiO_2.

For defining the Field Oxide region or for defining the Shallow Trench, we use Mask Number 1.

The first step is:

Choosing a P type substrate of orientation <100> and resistivity of 15Ω-cm.

Second step is :

Thermal Oxidation for 15 minutes at 900°C in steam ambient or 45 minutes at 1000°C in dry O_2 ambient. In either case we get 40nm Silicon Dioxide layer.

Third Step is:

Growing Silicon Nitride layer in a second oven at 800°C.

$$3SiH_4 + 4NH_3 = Si_3N_4 + 12H_2 \textit{ at low pressure hence it is LPCVD}$$

_____6.19

Fourth Step is:

We spin a Photo-Resist layer. Bake it at 100°C. Expose it through Mask 1. Develop it. Fix it so that the mask portion is resistant of the etchant. Next we do dry etching by Fluorine Plasma to etch out the Si_3N_4 . In the masked region Si_3N_4 is retained and elsewhere a window is made .

Fifth Step is:

Next fixed Photo-Resist layer is stripped by a stripper or stripped by Oxygen plasma.

Sixth Step is:

Next we do local oxidation (LOCOS) or shallow trench isolation(STI). The end result is shown in Figure 6.44 and Figure 6.45

Seventh Step:

By dry etching the remaining portion of Si_3N_4 layer is removed.

Eighth Step :

Mask 2 is used for retaining hardened KPR in PMOS and Boron is ion-implanted in NMOS region as shown in Figure 6.46. Implant dose will be 10^{13} per cm^2 at an energy of 150 to 200keV. This will ensure a P-tub of 5×10^{16} to 10^{17} per cm^3. By high temperature annealing the damage is healed. This may be 10sec at 1000°C or 30 min at 800°C. Hardened KPR is stripped by stripper or by Oxygen Plasma.

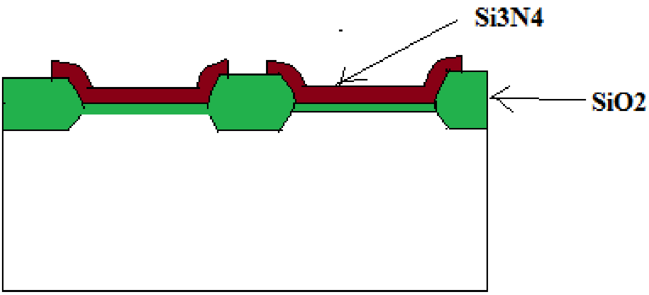


Figure 6.44 LOCOS for isolation but results in Bird's Beak

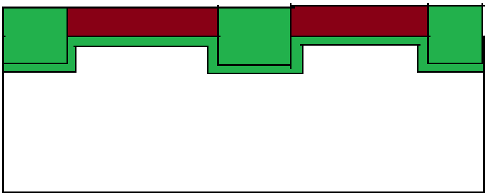


Figure 6.45 Shallow Trench Isolation(SIT) with no Bird's Beak

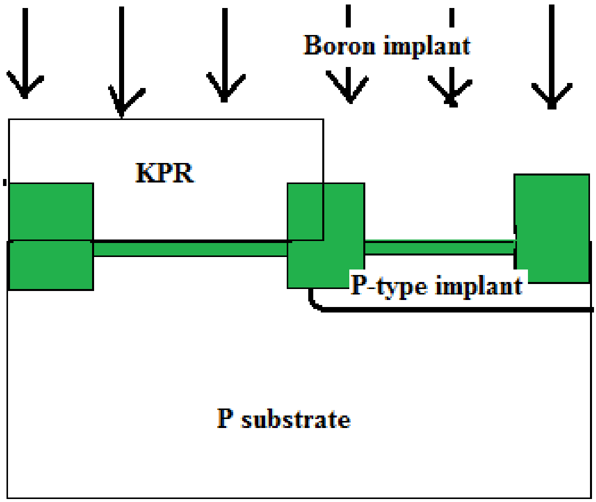


Figure 6.46 KPR is used for masking the region where PMOS will be built using Mask 2. A Boron implant rovides the doping for NMOS device.

Ninth Step:

We use Mask 3 to implant the N-tub as shown in Figure 6.47. Phosphorous is almost 3 times heavier than Boron hence Phosphorous ions will be implanted at 300 to 400 keV. This will ensure the same kind of doping as in P-tub.

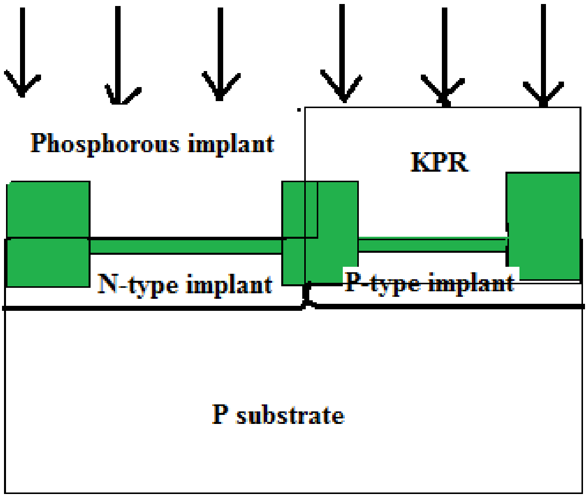


Figure 6.47 KPR is used to mask the regions where NMOS will be built using Mask 3. A Phosphorous implant provides the doping for N-tub for PMOS devices.

Tenth Step:

High temperature drive-in of P-tub and N-tub is done for 4 to 6 hours so as to drive the depth to 3 microns as shown in Figure 6.48.

There are several options for active regions and Tub formation. We will study only two options which are generally adopted in MOS fabrication technology. These are :

Option 1__Field implant. This provides a better isolation.

Option 2__Buried layer and Epitaxial layer. This minimizes the possibility of Latch-up through the parasitic BJTs.

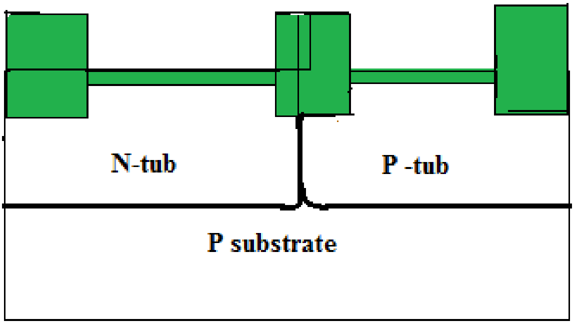


Figure 6.48 A high temperature drive-in completes the formation of N-tub and P-tub.

SSPD_Chapter 6_Part 6_MOS ICs Fabrication Technology_to be continued2

This is a continuation of MOS Fabrication technology. It describes the process option of P implant below the field oxide to make component isolation doubly sure and the process option of buried layer and epitaxial layer to prevent Latch-Up.

SSPD_Chapter 6_Part 6_MOS ICs Fabrication Technology_to be continued2

Section 6.6.2.4 Process Options for Active Region and Well Formation.

There are many options for Well Formation but here we will consider only two options which are commonly used by the Industries. First option is “Field Implants under LOCOS region”. This option is a more reliable method of lateral isolation of the active MOS. Second option is “Buried and Epitaxial Layer”. The second option prevents Latching of MOS through parasitic BJT formation.

Option 1. “Field implants under LOCOS region”.

In the CMOS process flow till now, the implant energies of P and N tubs were carefully chosen to penetrate the thick field oxide. Implants would be masked only in fixed KPR covered area. When the implant penetrates the field oxide, it increases the substrate doping underneath the LOCOS region and prevents the formation of inversion layer under normal voltage condition. Thus field oxide isolation is achieved. But the stopping by thick oxide is statistical in nature. Hence the dose which needs to be implanted underneath the field oxide is not achieved while the well implant is taking place. To ensure an implant under the field oxide this implant has to be carried out in a separate step.

In Section 6.4.3 , Figure 6.7, the middle diagram contains the windows for local oxidation. The active regions are shielded by KPR/Si_3N_4/SiO_2 .Now Boron ions of implant dose of $1\times10^{13}cm^{-2}$ at 50keV is used for field implant. The implant energy will make the Boron ions pass through thin SiO_2 region but will be completely stopped by KPR/Si_3N_4/SiO_2 stack. The LOCOS region acquires a shallow P-type implant as shown in Figure 6.49.

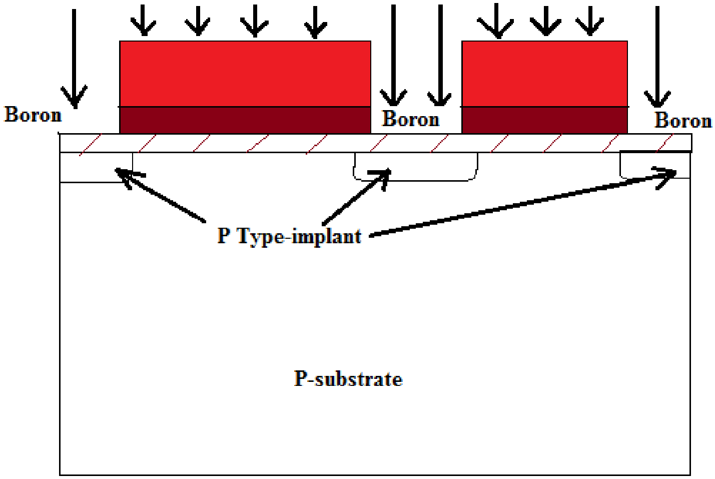


Figure 6.49 Process option for active region formation. Boron implant prior to LOCOS oxidation increases the substrate doping locally under the LOCOS region to minimize the field inversion problems.

After the P-type implant we remove the fixed KPR by oxygen plasma or by an organic solvent known as stripper. Next we do the Local Oxidation through the aperture in Si_3N_4. The structure obtained is shown in Figure 6.50.

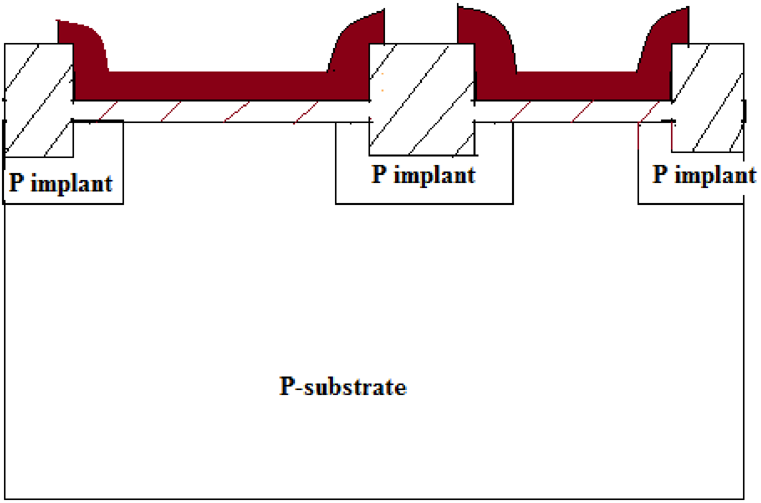


Figure 6.50 Process option for P-type implant under field-oxide.

By dry etching Si3N4 is removed. Now we follow the steps in Figure 6.46 and Figure 6.47. The final structure is shown in Figure 6.51.

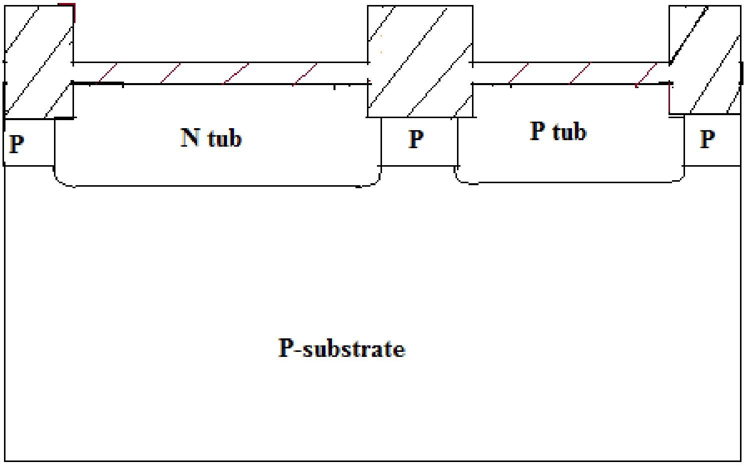


Figure 6.51 After N tub and P tub are implanted through thin oxide at low implant energies of 50keV.

Option 2. “Buried and Epitaxial layer” process option.

In CMOS formation there is always a problem of parasitic BJTs as shown in Figure 6.52

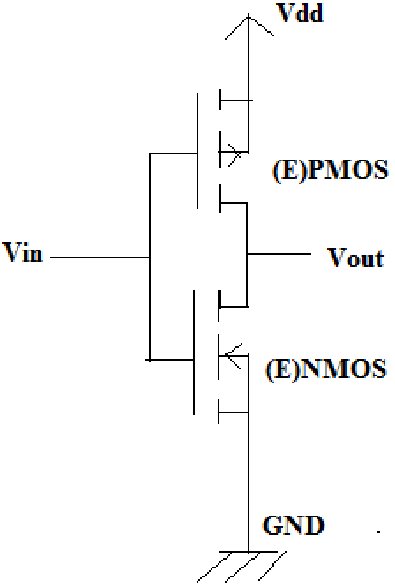


Figure 6.52.A. Circuit Configuration CMOS Inverter Logic Gate.

In the circuit given in Figure 6.52.A we find that body of (E)PMOS which is N-tub is connected to the source of PMOS which is P+.

Similarly body of (E)NMOS which is P-tub is connected to the source of NMOS which is N+.

Source to Bulk Voltage (Vsb) is kept zero in both cases so that Threshold voltage is maintained constant. If Vsb is not zero then Threshold Voltage will become a function of Vsb given by the following relation:

$$V_{TN} = V_{TO} + \gamma\sqrt{V_{SB} + 2\varphi_F} - \sqrt{2\varphi_F} \tag{6.20}$$

γ = body-effect parameter;________________________________

2φF = surface potential parameter;_____________________________

Since both MOSs are Enhancement type hence both are Normally OFF type therefore their channels are shown broken.

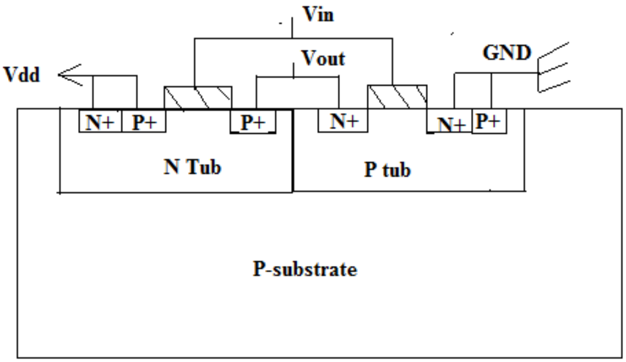


Figure 6.52.B. The cross-sectional view of a typical CMOS Inverter Logic Gate.
(E)PMOS source is connected to Vdd and is also connected to N tub through N+ to minimize the body effect of N tub;
(E)NMOS source is connected to the Ground and is also connected to P tub through P+ to minimize the body effect of P tub.

A non-zero value of source-bulk voltage(Vsb) causes the threshold voltage to change depending on the body-effect paameter. To maintain a constant threshold voltage we give the option of maintaining Vsb = 0.

The P+ source of PMOS in N-tub and N+ source of NMOS in P-Tub together with N-Tub bulk and P-Tub bulk behave as cross-coupled BJTs.

P+ source of PMOS in N-tub, N-Tub and P-Tub behave like a parasitic PNP BJT where P+ is the emitter.

N+ source of NMOS in P-Tub. P-Tub and N-Tub behave like a parasitic NPN BJT where N+ is the emitter.

The cross coupling of the two gives rise to a four-layer device namely SCR as shown in Figure 52.C. If this SCR fires, because of the voltage drop in N-tub bulk resistance and because of voltage drop in P-tub bulk resistance, then it will provide a low impedance path from Positive Vdd supply to the ground leading to irreversible destruction of the CMOS structure. To prevent this from happening we use the second option. The active devices are formed in lightly doped epitaxial layer. This epitaxial layer is deposited over buried layer. N+ buried layer shorts N-tub bulk resistance and P+ buried layer shorts P-tub bulk resistance. This prevents the firing of SCR.

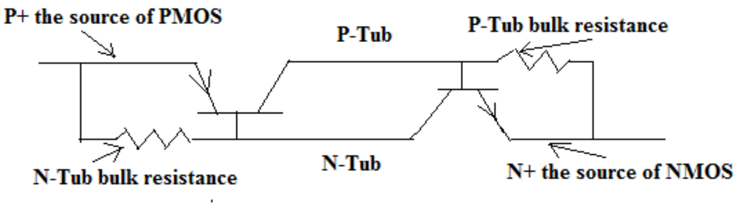


Figure 6.52.C. PMOS in N-Tub and NMOS in P-Tub give rise to two cross coupled parasitic BJTs which acts as a four layer SCR. Because of high N-Tub bulk resistance and high P-Tub resistance both BJTs can turn-on leading to the firing of SCR which will lead to huge current dran from Vdd to Ground and in the process CMOS will be destroyed.

The process steps for the second option.

We start from Step 1 as shown in Figure 6.53. On P-Type substrate SiO2 , 40nm thick, is grown. That is followed by Si3N4, 80nm thick.. Lastly it is covered by 200nm thick KPR layer.

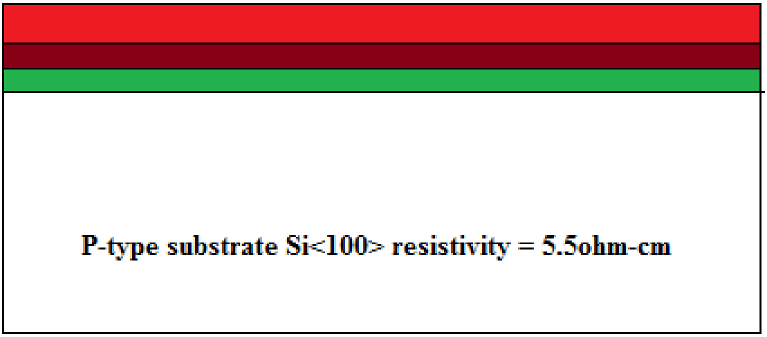


Figure 6.53. P-Type Wafer has 40nm thick SiO2 layer in green, 80nm thick Si3N4 layer in brown and 200nm KPR layer in red.

Mask 1 defines the buried layer area namely N+ type buried layer as shown in Figure 6.54. We want N-type buried layer with high doping concentration. So we want a donor type dopent which does not out-diffuse in subsequent thermal cycles. Antimony and Arsenic both qualify for buried layer diffusion but Arsenic has higher solid solubility therefore we go for Arsenic in a high dose implant of $10^{15}cm^{-2}$ at accelerating voltage of 50kV. This energy of 50keV is sufficient for implanting it beneath the 40nm thick Oxide layer.

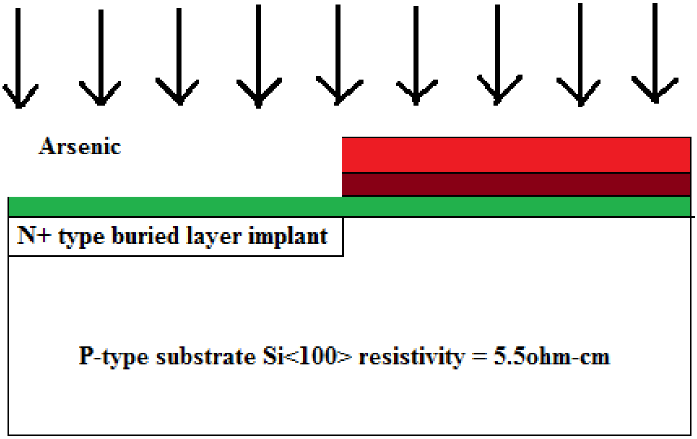

Arsenic
N+ type buried layer implant
P-type substrate Si<100> resistivity = 5.5ohm-cm

Figure 6.54.Process option incorporationg buried and epitaxial layer . Mask 1 defines the the region for N+ buried layers. Arsenic dopent is used to achieve N+ buried layer in the area defined by Mask 1. Arsenic implant is used because Arsenic has negligible diffusion coefficient hence it undergoes minimum out-diffusion in subsequent heat cycles.

After N+ type buried layer diffusion, we strip off the fixed KPR and carry out drive-in for 2 hours at 1000°C, with 60 minutes of this in steam ambient for LOCOS oxidation. Due to drive-in the buried layer is driven to depth of 2microns and just over the buried layer we get a local thick layer of SiO_2 400nm thick. The cross-sectional view at this point is shown in Figure 6.55.

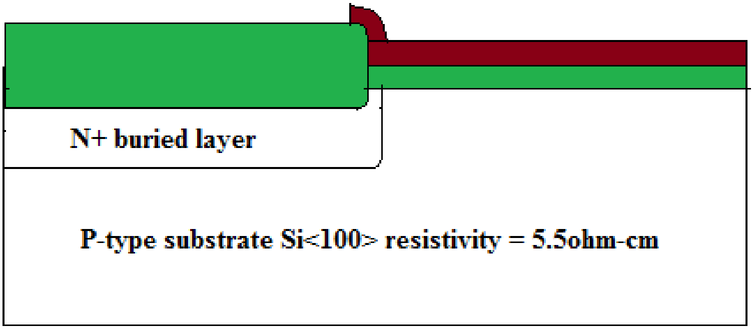


Figure 6.55 The N+ buried layer is driven in partially in steam ambient. LOCOS oxide forms only above the N+ region.

This high temperature drive-in plus LOCOS oxidation accomplishes the following objectives:

1. It drives in the buried layer to 2 micron depth;
2. Thick oxide layer of 400nm is achieved just over the buried N+ layer because of the masking effect of Si3N4 layer;
3. This LOCOS oxide allows the same mask to be used for defining both N+ buried layer as well as P+ buried layer and the LOCOS oxide allows self alignment between the two buried layers;
4. Finally this step creates a step in the silicon surface at the edges of the N+ buried layer because oxidation consumes silicon. This step will act as the foot print of the buried layer in later processing.

After the furnace operation, Si3N4 is dry etched in Fluorine Plasma Chamber. Now we do the ion implantation of P+ buried layer. The thick oxide layer (400nm thick) over N+ buried layer acts as the shield against P+ Boron ion implant but thin oxide layer (40nm thick) elsewhere allows the P+ Boron implant to enter the silicon surface. An implant dose of $10^{14}cm^{-2}$ at accelerating voltage of 50-75 kV. The implant for P+ buried layer is Boron. A lower implant dose is chosen because Boron has a higher diffusion coefficient and to prevent out-diffusion in subsequent thermal cycles a lower implant dose helps. The cross sectional view at this point is as shown in Figure 6.56.

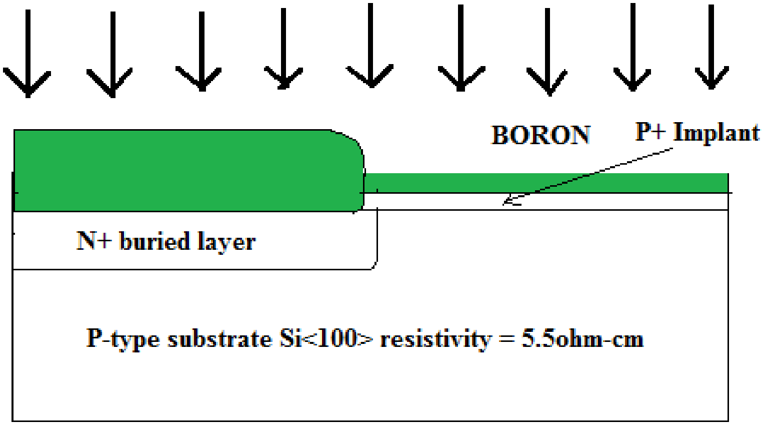


Figure 6.56 The P+ buried layer is implanted using LOCOS oxide as a mask to shield N+ buried layer.

A simple drive-in cycle is carried out for several hours at 1000-1100°C in inert ambient of Nitrogen or Argon. LOCOS oxidation is not required. The resulting structure is shown in Figure 6.57.

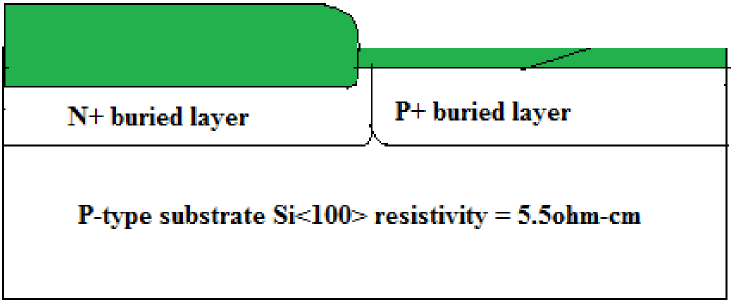


Figure 6.57. The self aligned N+ and P+ buried layers after the P+ drive-in cycle. During this drive-in cycle, N+ buried layer is also driven-in but only marginally because of much lower diffusion coefficient of Arsenic.

Now we require lightly doped N-tub and P-tub. Lightly doped N-tub and lightly doped P-tub are required for fabricating PMOS and NMOS respectively because of the following reasons:

1. The light doping reduces the junction capacitances hence improves the frequency response;
2. It gives a better control and a lower threshold voltage or turn on voltage.

Threshold voltage expression is given in Equation 6.21.

$$V_{Th} = V_{FB} + 2\varphi_F + \frac{\sqrt{2\epsilon_S q N_A (2\varphi_F)}}{C_{OX}}$$

--6.21

Heavy doping increases the surface potential and hence increases the turn-on voltage.

Now this lightly doped region cannot be achieved by counter doping and by compensation because the buried layer is of the order of 10^{19}/cc and active regions have to be of 10^{16}/cc. Therefore we choose the option of lightly doped epi-layer.

We etch out the SiO2 layer completely and grow intrinsic epitaxial layer of Silicon to a thickness of 2 to 3 microns. The cross sectional view is as shown in Figure 6.58. The step at the boundary of the buried layer has propagated upward into the epitaxial layer also and it will help in subsequent alignment of the masks.

After this the steps carried out in Figure 6.44 or in Figure 6.45 to grow the field oxide and the steps carried out Figure 6.46 , in Figure 6.47 and in Figure 6.48 to grow the twin tubs and drive them in are carried out. The drive-in for both the tubs are carried out in a way that the twin-tub are linked up with one-another and the buried layers are also linked up. The final structure appears as shown in Figure 6.59.

It may be noted here that this option of buried layer plus epitaxy has required only one more Mask but many more steps.

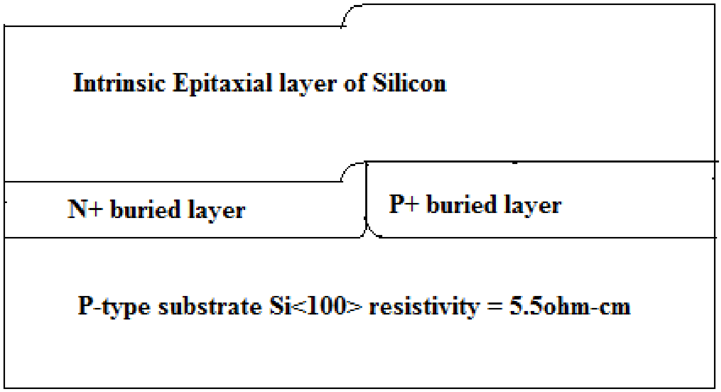

Intrinsic Epitaxial layer of Silicon
N+ buried layer
P+ buried layer
P-type substrate Si<100> resistivity = 5.5ohm-cm

Figure 6.58 SiO2 is completely stripped of and 2 to 3 micron thick Si epitaxial layer is grown by CVD. The boundary of the buried layer is seen as a step in the epitaxial layer. This step is visible under microscope and helps in subsequent alignment of the mask.

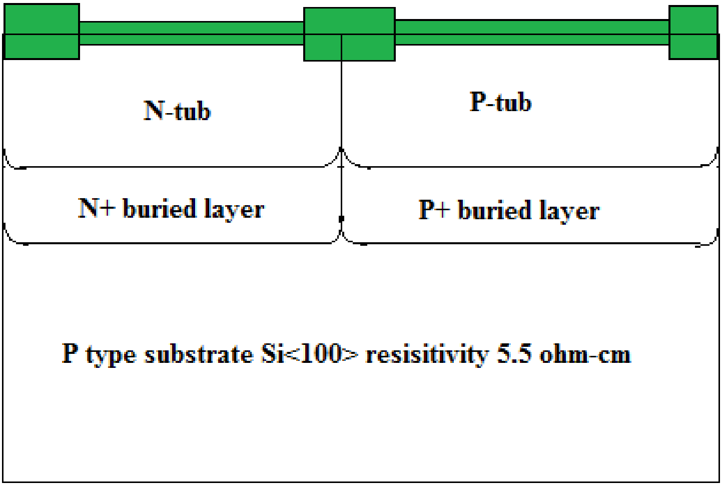


Figure 6.59 Process Option incorporating buried layer and epitaxial layer.

Now we are ready for the final fabrication of complementary symmetry MOS devices.

SSPD_Chapter 6_Part 6_MOS IC's Fabrication Technology_To be continued3

SSPD_Chapter 6_Part 6_to be continued3 describes CMOS flow process after the twin tub has been realized. This part describes how Lightly Doped Drain is formed,sidewall spacers are formed and finally the Source and Drain are formed in the active region of the device.

SSPD_Chapter 6_Part 6_MOS Fabrication Technology_Concluded

Section 6.6.3 Gate Formation.

In the preceding section we saw how the twin tub is fabricated for realizing the complementary symmetry MOS also known as CMOS. If we choose the process option of P-implant below Field Oxide then the crss-sectional view is as shown in Figure 6.51. If we choose the process option of incorporating buried and epitaxial layer then the cross-section looks like Figure 6.59. For simplicity we will start with the twin tub as described in Figure 6.48. Now the critical part of MOS has to be fabricated namely P-channel and N-channel in normally-off PMOS and normally-off NMOS respectively. In other words we are fabricating (E)PMOS and (E)NMOS.

For the convenience of the reader Figure 6.48 is being reproduced here as Figure 6.60.

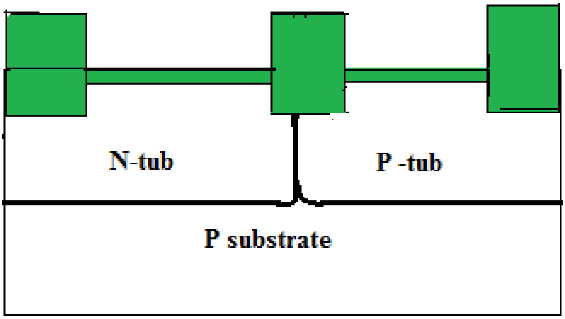


Figure 6.60 The starting point of the realization of (E)PMOS in N-tub and (E)NMOS in P-tub. Green colour is SiO2.

6.6.3.1. To precisely control the threshold voltage in the range 0.5 to 0.8 Volt, we do a second ion-implant of Arsenic or Phosphorous in N-tub and Boron in P-tub sequentially just below the gate oxide. The dose of this second ion-implant is calculated from Equation 6.22.

$$V_{Th} = V_{FB} + 2\varphi_F + \frac{\sqrt{2\epsilon_S q N_A(2\varphi_F)}}{C_{OX}} + \frac{qQ_I}{C_{OX}}$$

---6.22

Q_I = implant dose in atoms per cm^2_______________________________

Here V_{FB} = flat band voltage = gate voltage required to compensate for the work function difference in the gate metal and semiconductor = $\varphi_M - \varphi_S$;________________________

N_A = doping density of P-tub which is going to host (E)NMOS before the critical ion-implant of Boron;________________________________

And ε_S = permittivity of Silicon = absolute permittivity of free space × relative permittivity of Silicon;______________________________

C_{OX} = unit area capacitance of the gate oxide;

$$C_{ox} = \frac{\epsilon_S}{t_{ox}}$$

t_{OX} = thickness of the gate oxide . This has to be precisely controlled inorder to achieve a precision in the threshold voltage.

Here it is assumed that implant dose is located near the Si-SiO_2 interface inside the respective MOS depletion region.

Mask 4 is used for selectively ion implant the P-tub with Boron at a dose implant of 1 to 5 × 10^{12}/cm^2 at 50 to 75 keV. The energy is high enough to pass through the gate oxide but low enough to lie near the Si-SiO_2 interface,

Mask 5 is used for selectively ion implant the N-tub with Arsenic or Phosphorous at a dose implant of 1 to 5 × 10^{12}/cm^2 at 100 keV. The P-channel implant in N-Tub would be at a higher energy because

Arsenic is heavier.

So the final cross-sectional view would be as shown as in Figure 6.61.

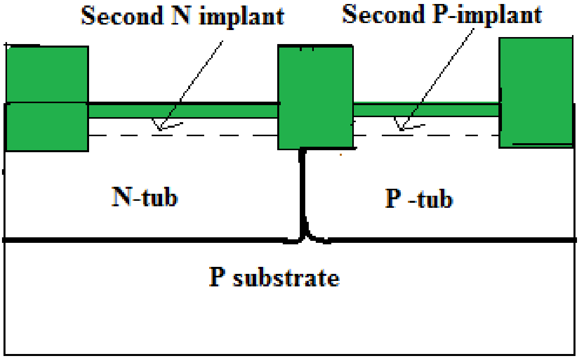


Figure 6.61. The cross-sectional view after the second implant to precisely control the threshold voltage of the complementary symmetry MOS's

6.6.3.2. Regrowth of Gate Oxide to the precision required for the Threshold Voltage.

Gate area Silicon Dioxide layer has undergone two times battering under the two ion-implants hence the quality has deteriorated. Therefore it is important that it is stripped by dilute hydrofluoric acid and regrown to a precise thickness as demanded by the threshold voltage. Table 6.13 gives the typical gate oxide thickness for different generations of technology. After the gate-oxide is regrown to the precise thickness we go to the step of Poly-Si(heavily doped)Gate contacts. PolySi gate contacts were used before 45nm for achieving lower threshold voltage. But at 45nm technology, leakage current from the gate and the subthreshold current through the channel started deteriorating leading to increased thermal dissipation even in stand-by condition therefore thicker gate oxide thickness along with high – K dielectric such as Silicon Oxynitride was adopted as the gate dielectric. This reduced the standby dissipation but created the new problem of Fermi-pinning which led to the loss of control over the threshold voltage. Hence PolySi was replaced by Metal Contact. Thus at 45nm and below generation of technology, PolySi-SiO2 was replaced by MG-HK.

Table 6.13. Gate Oxide Thickness vs Channel Length

for different generations of Technology.

L(micron) (Poly-Si)	0.35	0.25	0.18	0.13	0.10	0.07	45nm (MG-HK)	32	22	15
Tox(nm)	8	6	4.5	3.4	2.5	1.9	4	3.4	2.5	1.2
Ioff(nA/micron)		0.5	3	10	40	100	200	300	300	300
Fringing Capacitance_Cf*(fF/µm)		0.60	0.71	0.61	0.74	0.78	0.74	0.85	1.04	1.01

Transistor delay_τ(ps)	6.26	4.56	2.55	1.88	1.46	1.15	1.23	1.47	1.45
Sheet resistance of heavily doped PolySi(ohms per sq)	300	250	200	200	200	200	200	200	200

Table 6.13 is based on projection and simulation for nodes at 45nm and below. For nodes above 45nm, the Table is based on actual experimental work. It is clear that at 45nm and at greater scaling, we get counterscaling effects. Because of increased fringing capacitance, transit delay deteriorates beyond 45nm technology. But with emergence of multicore processors, number of available transistors is more important than making them faster. Hence scaling will continue to be used even though it is becoming counter productive.

6.6.3.3 LPCVD of PolySi for Gate contacts.

By Low-Pressure-Chemical-Vapour-Deposition techniques Si is deposited over the entire area of SiO_2 as shown in Figure 6.62. Silane is passed over the wafer and at 600°C it decomposes into Si and H_2. Si precipitates over SiO_2 surface. Since the substrate of SiO_2 is amorphous hence LPCVD Si is

PolySi.

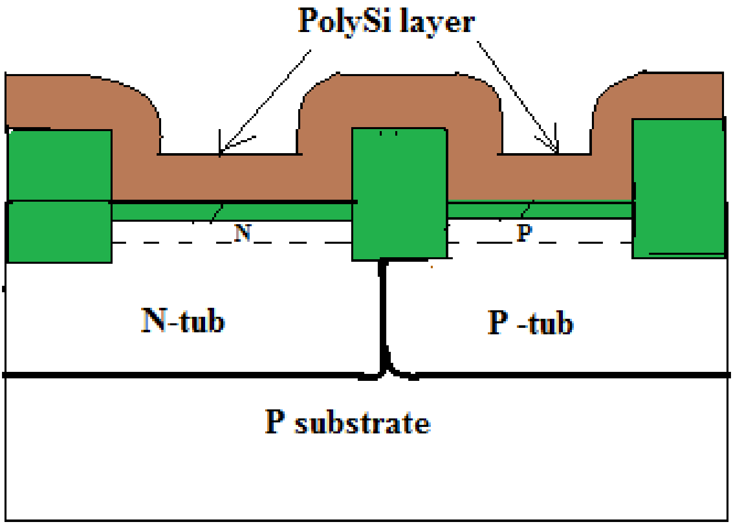


Figure 6.62. LPCVD of PolySi over the entire wafer. This is followed by ion-implantation of Donor atoms.

300 nm to 500 nm thick PolySi is deposited over the entire wafer area. By unmasked ion-implant of Arsenic or Phosphorous , PolySi is heavily doped provideng sheet resistance of 300 ohms per square. The N-type ion-implant has a dose of $10^{15}/cm^2$ at an energy of less than 50keV so that it implants the PolySi but does not penetrate through to the underlying gate oxide. In subsequent thermal cycles N-type donor dopent will diffuse and evenly distribute throughout the PolySi contact.

In some PolySi deposition systems, in-situ doping is done while LPCVD is carried out. In this case ion-implant step of heavy doping is not required.

6.6.3.4. Formation of PolySi Gate contacts in MOS active areas.

Using KPR and Mask 6, the PolySi is retained in Gate areas of MOS and PolySi is etched out from the remaining portions of the wafer as shown in Figure 6.63. This again is dry plasma etching using chlorine or bromine based plasma.

Heavily doped PolySi is not only used as Gate contacts but also as local interconnects between NMOS and PMOS gate contacts but these cannot be used for long interconnects like metal interconnections. PolySi has relatively high sheet resistance (300Ω/square) as compared to that of Metal interconnects which is only 0.1Ω/square. The RC delays of long poly lines make these long poly lines unacceptable in IC fabrication.

While etching PolySi we want selectivity as well as anisotropy. Selectively will ensure that only PolySi is etched and not the underlying SiO2 and anisotropy will ensure that there is no undercutting. Now these are two conflicting requirements. There has to be a trade off between the two in our plasma etching systyems.

The remaining fixed KPR is removed by stripper solution or by oxygen plasma etching.

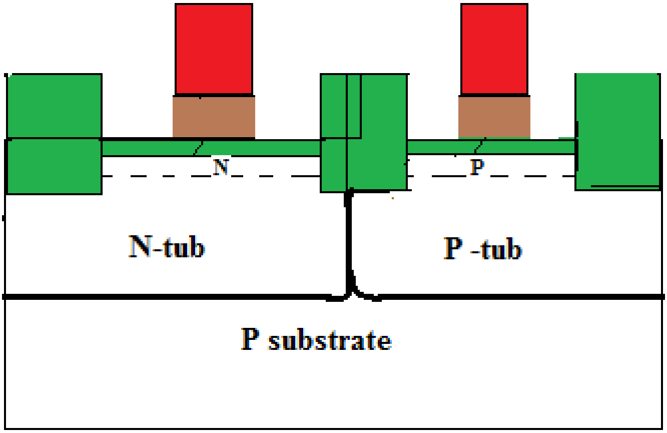


Figure 6.63.KPR is applied and mask 6 is utilized to define the regions where MOS gates are located. PolySi is etched out using plasma etching from unwanted area.

6.6.3.5. Self-aligned Source and Drain fabrication.

Using mask 7 light doping implant of Phosphorous is done in NMOS area situated in P-tub as shown in Figure 6.64. Typically a dose of 5×10^{13} to 5×10^{14} per cm^2 at 50keV is used.

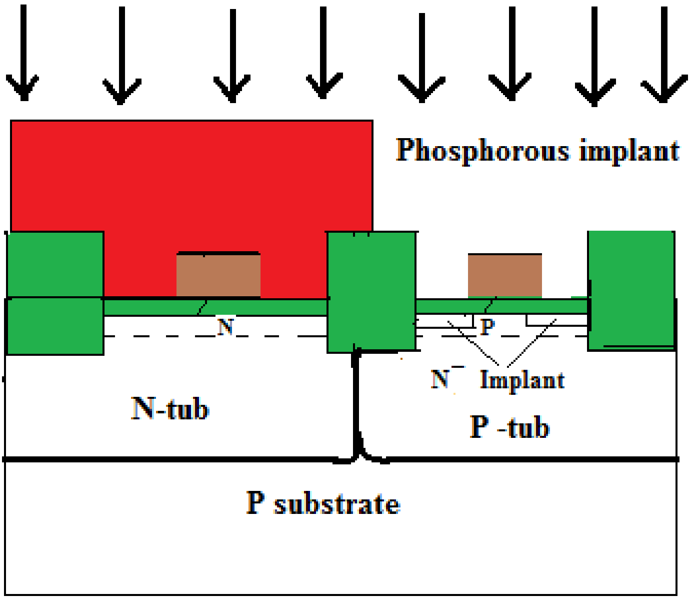


Figure 6.64. Mask 7 is used to cover PMOS device in N-tub. A light dose Phosphorous implant is used to form the the tip or extension (Light Doped Drain-LDD) regions in NMOS device in P-tub.

Using mask 8 light doping implant of Boron is done in PMOS area situated in N-tub as shown in Figure 6.65. Typically a dose of 5×10^{13} to 5×10^{14} per cm^2 at 50keV is used.

Once the light dose implant is completed, fixed KPR is completely removed from the surface of the wafer by stripper or oxygen plasma.

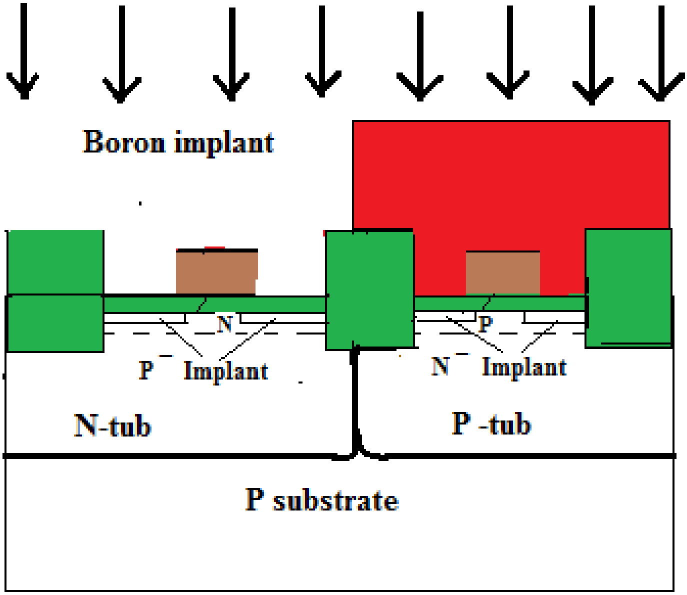


Figure 6.65. Mask 8 is used to cover NMOS device in P-tub. A light dose Boron implant is done to form the tip or extension regions(LDD) in PMOS device situated in N-tub.

Note that in Figure 6.64 and in Figure 6.65, LDD implant lies under the edge of the gate. This has been achieved by tilted angle implant.

6.6.3.5.1 Sidewall Spacers along the edge of the PoySi gate.

Now we do LPCVD of a conformal spacer dielectric layer (SiO2 or Si3N4) on the wafer surface. The thickness of this layer will decide the width of the sidewall spacer which is chosen around 100nm to optimize the device characterisics.

At 400°C, $SiH_4 + O_2 \rightarrow SiO_2 + 2H_2$;

Or at 900°C, $SiH_2Cl_2 + 2N_2O \rightarrow SiO_2 + 2N_2 + 2HCl$;

Conformal spacer dielectric layer deposition is obtained as shown in Figure 6.66. Note that the thickness of SiO_2 is much thicker along the edges of the polysilicon as compared to that on the flat regions of the wafer surface. This thickening is caused due to conformal deposition on the vertical edge of the PolySi gate contact.

Now we use fluorine-based plasma to carry out highly anisotropic etching of SiO_2 over the entire wafer area(etches vertically but not horizontally).

Deposited conformal layer of SiO_2 is removed everywhere except at the vertical edges of PolySi gate contacts. Thus sidewall spacers are realized as shown in Figure 6.67. Note that SiO_2 has been removed from Source and Drain region also as a part of the Sidewall Spacer formation process.

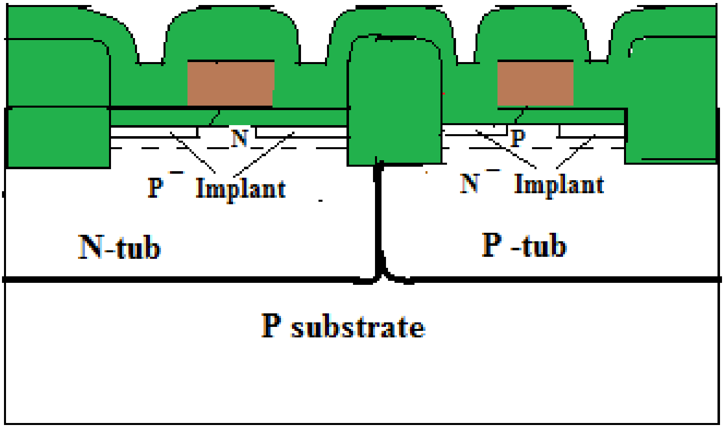


Figure 6.66. Conformal deposition of SiO2 on the wafer surface in preparation for side-wall spacer formation.

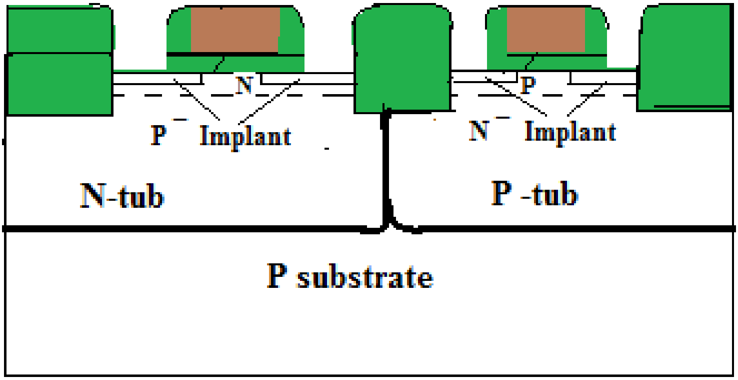


Figure 6.67. The deposited SiO2 is etched back anisotropically leaving sidewall spacers along the edges of PolySi gate contacts.

6.6.3.5.2. Final Source/Drain formation.

In the process of Sidewall Spacer formation, the oxide layer has been completely removed from Source and Drain regions also. So another 10nm SiO_2 is thermally grown over the whole wafer area. This provides a 10nm ‘screen oxide’ in the source and drain regions as well as on top of the PolySi gate regions.

The ‘screen oxide’ is needed for the following reasons:

1. Oxide being amorphous it randomizes the implanted ions trajectory thereby preventing Chanelling Effect. The Silicon lattice is crystalline and if the ion velocity vector is aligned with the crystal structure of Silicon, ions can go down the “channels” between lattice centers for long distances without encountering Silicon atoms. If this happens, the range of implanted ions will be much larger than expected. This is undesirable hence ‘screen oxide’ is needed.
2. ‘screen oxide’ helps minimize the incorporation of trace impurities into the Silicon from the implanter system.

After ‘screen oxide’ has formed we use Mask 9, to selectively implant P-tub with N+ source and N+drain. The N+ implant is Arsenic because in modern devices we need shallow junctions in small geometry devices. An implant of high dose of 2×10^{15} to 4×10^{15} per cm^2 at 75keV is used. This allows the implanted ions to cross ‘screen oxide’ but still be masked by fixed KPR. This completes the step of formation of (E)NMOS as shown in Figure 6.68.

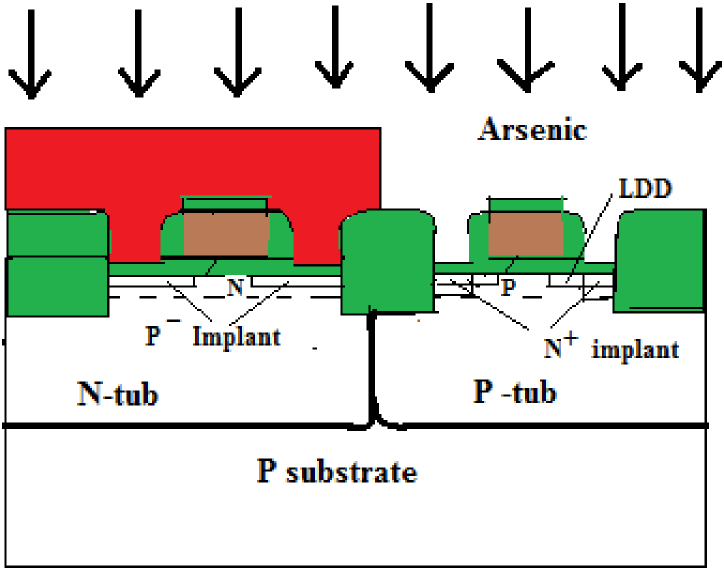


Figure 6.68. After growing thin 'screen oxide',KPR is applied and Mask 9 is used to protect PMOS regions. An Arsenic implant then forms he final Source and Drain of NMOS in P-tub.

Now Mask 10 is used to implant P+ Source and Drain of (E)PMOS in N-tub. For this using Mask 10 N-tub regions are kept exposed through appropriate windows in KPR layer. After developing the KPR layer, the existing KPR is hardened and is fixed in the oven so that it can shield the implanted ions from entering the shielded areas. The implant is high dose of Boron at less than 75keV. The dose itself is 1×10^{15} to 3×10^{15} per cm^2. Since Boron is lighter therefore it requires less energy to reach the same range.

In both these cases high dose implants have been used to minimize the parasitic resistances associated with the drain and sources. The final structure after the use of Mask 9 and 10 and after the final high temperature drive-in is as shown in Figure 6.69.

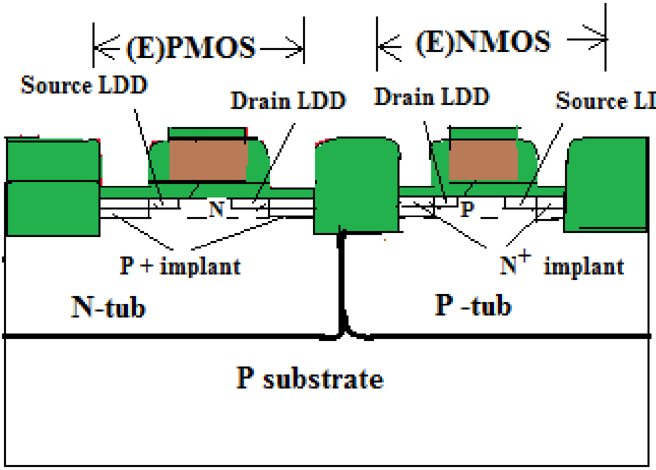


Figure 6.69. The cross-sectional view of the wafer after the final Mask step has been implemented and final drive-in carried out to activate all the implaned dopants and diffuse junction to their final depths.

6.6.4. Why do we have LDD and sidewall spacers along the edges of PolySi Gate Contacts?

In VLSI era as we move to deep-sub-micron and ultra-deep-sub-micron devices we are forced to adopt small geometry and shallow junction devices. At these levels of miniaturization many new effects have shown up and which have to be taken care of. In Figure 6.70 we give the magnified view of Figure 6.69.

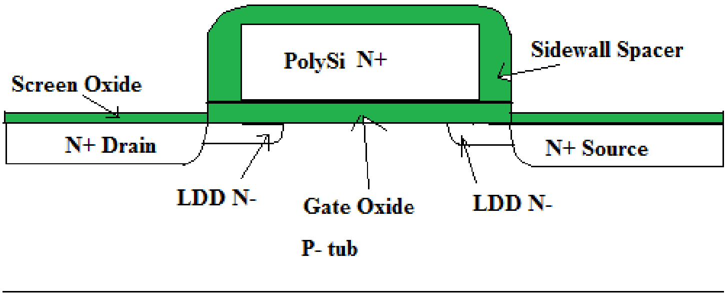


Figure 6.70. The detailed structure of deep-sub-micron (E)NMOS with LDD and Sidewall Spacer along the PolySi Gate Contacts.

Because of the scaling trends and its concomitant effects we have been forced to introduce Lightly Doped Drain(LDD) or Lightly Doped Source and Sidewall Spacer along the edges of PolySi(N+) Gate Contact.

As device dimensions have been scaled the operating voltage has not been proportionately scaled as a result the average electric field has increased by one order of magnitude. This creates “hot electron” effects. The hot electron experiences “scatter limited velocity”. Here the drift velocity is not proportional to the electric field hence electron drift velocity in-effect saturates.

Plus “hot electron” cause carrier injection into gate oxide by overcoming the Si/Silicon Oxide interface potential barrier of 3.2eV. These carriers trapped in oxide will cause instability.

By grading the doping of Drain and Source from N+ to N- Lightly Doped Drain(LDD) , we allow the drain voltage to drop over larger distances leading to reduction in the peak of the electric field caused from Drain To Source. Even modest reduction in the peak electric field leads to significant improvement in the reliability of the device.

At deep-sub-micron level, short channel effect show up in a big way. Theoretically in Pentode Region, the Drain Current should be a function of Gate Voltage only but due to channel length modulation Drain Current is a function of Drain Voltage also as shown in Equation 6.23.

$$I_{DS} = \frac{K_n'}{2} \times \frac{W}{L} \times (V_{GS} - V_{TN})^2 \times (1 + \lambda V_{DS})$$

6.23

Where λ is channel length parameter it varies from 0.001 to 0.1 per volt.

As device geometry is scaled, this channel length parameter increases. By use of shallow junction this parameter can be minimized. LDD provides this shallow junction. To achieve LDD we use light dose implants along with sidewall spacer.

This sums up the reasons why LDD and sidewall spacer is used along the edge of PolySi Gate contact.

Conclusion of CMOS Flow Process
SSPD_Chapter6_Part6_concluded gives the back end processing of an IC chip. This describes the multilevel metal inteconnects for completing the circuit connections of the active devices and encapsulation of the finished IC with passivating layer for the final packaging.

SSPD_Chapter 6_Part 6_MOS IC Fabrication_concluded

6.6.5. Contact and Local Interconnect Formation.

A means of interconnecting the components on the wafer has to be found out and a means of bringing the input and output pads of the Silicon Chip to the external pins of the IC package has to be devised. At the present state of art of IC Technology, three levels of interconnections are provided. Lowest level is called Local Interconnects.

The first step in this direction is to remove the oxide from the contact pads of Source, Drain and Gate. No mask is required for this step. Just a short dip in HF acid removes the screen oxide, 10nm thick. This step is shown in Figure 6.71.

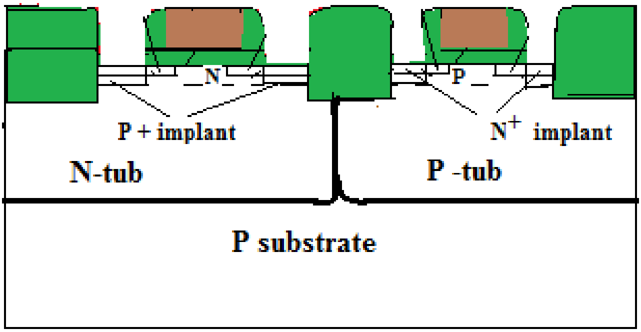


Figure 6.71 An unmasked oxide etch in dilute HF2 solution removes the SiO2 from the device source, drain regions and from the top surface of the Poly-Si gate.

Next the whole wafer is covered with Titanium by sputtering as shown in Figure 6.72. The metal layer is shown n black. The thickness of Ti is 50 to 100nm. Sputtering is one of the Physical Vapour Depositon technique. In a sputtering system, the desired material (Ti in this case) are physically knocked off a solid target with Ar+ ions. The Ti atoms are uniformly deposited. We will give the details of Physical Vapour Depositon in the next section.

Next the wafers are heated in N_2 ambient at 600 to 700°c for about 1 minute. There two kinds of reactions:

First is $Ti + 2Si \rightarrow TiSi_2$ where Ti is in contact with Si or Poly-Si. Titanium Salicide ($TiSi_2$) is an excellent conductor and shown in navy blue in

Figure 6.73. $TiSi_2$ forms the very low resistance contact pads to source, drain and gate.

Second is Ti + N→TiN where there is no contact with Si. . It is shown as light blue in Figure 6.73. TiN is also a conductor but not as good as metal hence used as ‘local’ interconnect. Long interconnects of TiN will cause unacceptable RC delays.

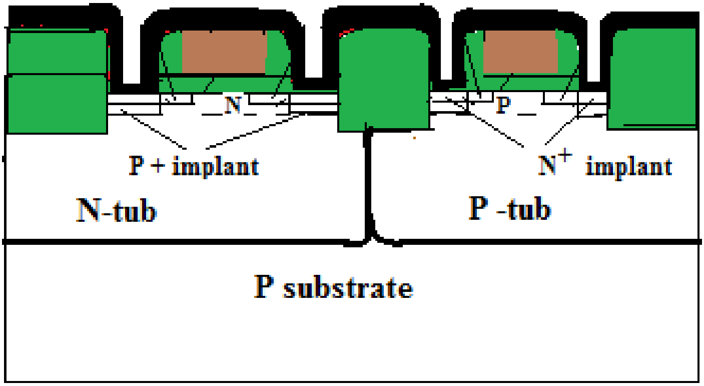


Figure 6.72 Titanium is deposited over the entire wafer surface by sputtering.

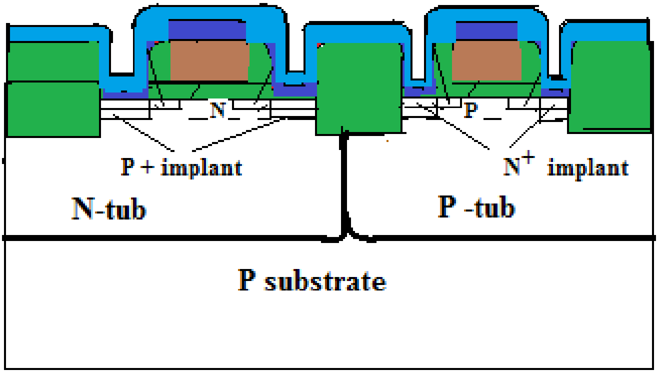


Figure 6.73 The Titanium is reacted in anN2 ambient, forming Titanium Silicide TiSi2 where it contacts Si or Poly-Si (navy blue regions in the figure) and TiN elsewhere (sky blue).

Mask 11 is used for etching out the local interconnect pattern as shown in Figure 6.74. Where SiN_2 is to be retained is protected by 200nm thick exposed KPR. Elsewhere KPR is removed by developer. Exposed KPR is annealed so as to become 100% resistant of the etchant. Next the wafer is dipped in Ammonium Hydroxide:Hydrogen Peroxide: Water solution in mass ratio of 1:1:5. This etchant etches away the exposed SiN. Thus we get the desired local interconnect pattern as shown in Figure 6.75. Hardened KPR is stripped away either by organic solvent or by O_2 plasma etching. After KPR stripping, we heat it in an inert ambient at 800°C for 1 minute. This reduces the sheet resistance of TiN and $TiSi_2$ to the final value of 10Ω/sq and 1Ω/sq.

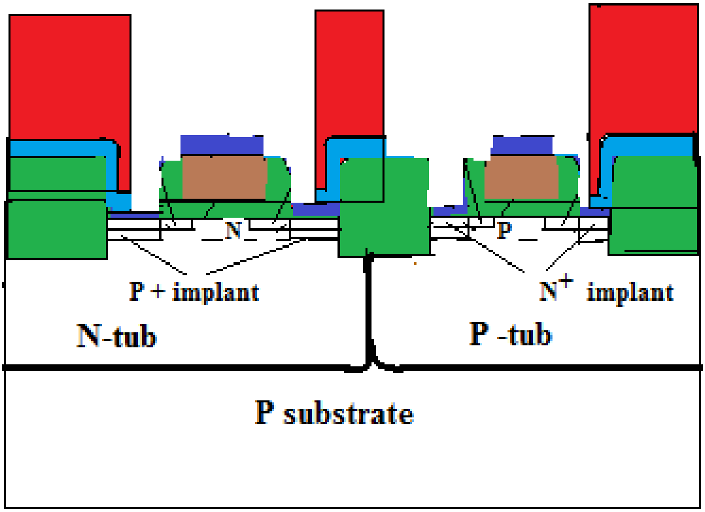


Figure 6.74 KPR is aplied and Mask 11 is used to define the regions where TiN will be retained to provide Local Interconnects. From the remaining regions TiN is etched away.

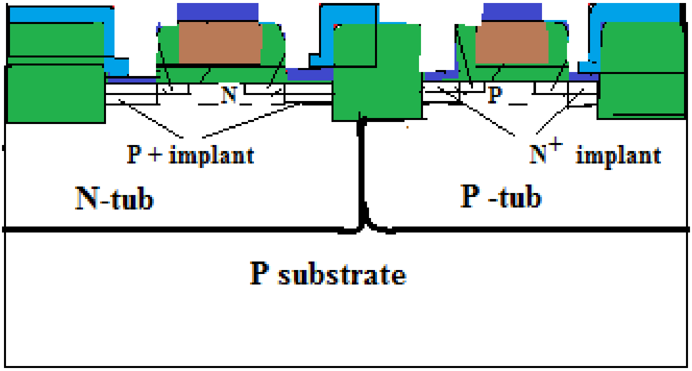


Figure 6.75 The desired LOCAL INTERCONNECT pattern after stripping the hardened KPR.

6.6.6. Multilevel Metal Interconnect Formation-The Back-End Process.

After making the local interconnect we have to make the second level interconnect(metal level1) and third level interconnect (metal level2). For this we do confomal deposition of SiO_2 over the portion shown in Figure 6.76. This is done by CVD or LPCVD. This is thick layer of SiO_2 1 micron thick. This helps cover up the undulations on the local interconnect level. If these undulations or non-planar nature of the first level interconnect is not evened out then it is going to give rise to serious reliability problem in the subsequent interconnections. This thick layer of conformally deposited Oxide is doped with Phosphorous or with Phosphorous + Born combined resulting in PhosphoSilicate Glass(PSG) and BoroPhosphoSilicate(BPSG) respectively. In some cases an undoped SiO_2 layer is added on top of PSG or BPSG.

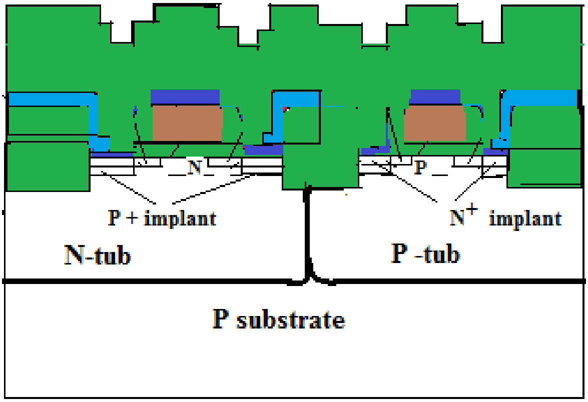


Figure 6.76. Conformal deposition of a thick layer of SiO2, 1 micron thick, over the entire wafer. The upper surface of this conformal deposition still has some undulations as the lower level of local interconnect.

Phosphorous has been added to protect the active devices from Na+ ions and Boron has been added so that the glass can reflow at a lower temperature. This heat cycle and reflow is required to even out the undulations after conformal deposition. But even after this reflow the undulations are not completely removed and planarization problem remains.

Planarization is done either by resist etchback technique or by CMP method. Resist etchback technique is as follows:

In resist etchback technique we spin KPR over the entire wafer surface creating a plane KPR surface at the top. All the hills and valleys are filled and a horizontal level of KPR is left at the top. Remember water seeks its own level. Now we can find a Oxygen Plasma etching combination which etches SiO_2 and KPR with equal ease. Allow this physical etching to continue until the lowest level in SiO_2 is reached. At this point stop the etching. In effect we have achieved complete planarization as shown in Figure 6.77.

In CMP method the wafer is put face down on a polishing machine and it is polished flat in a high-pH silica slurry. This also results in same planar surface as shown in Figure 6.77.

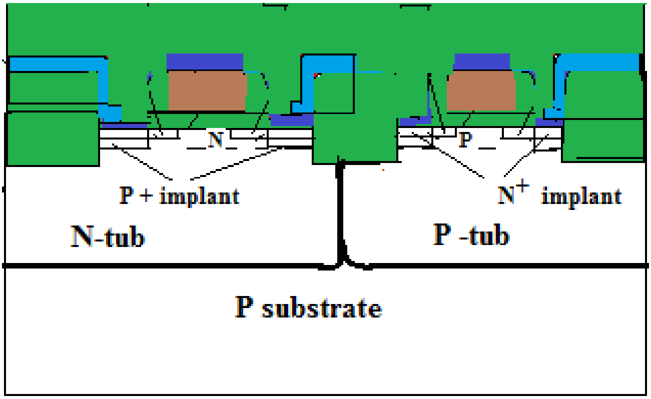


Figure 6.77 Chemical Mechanical Polishing or resist etchback is used to polish or etchback the deposited oxide layer. This planarizes the wafer surface.

6.6.6.1. Vias between Metal Level_1 and contact pads of the Active Devices.

We use Mask 12 to define the vias through which contacts will be made between Metal Lavel_1 and contact pads of the Active Devices as shown in Figure 6.78.

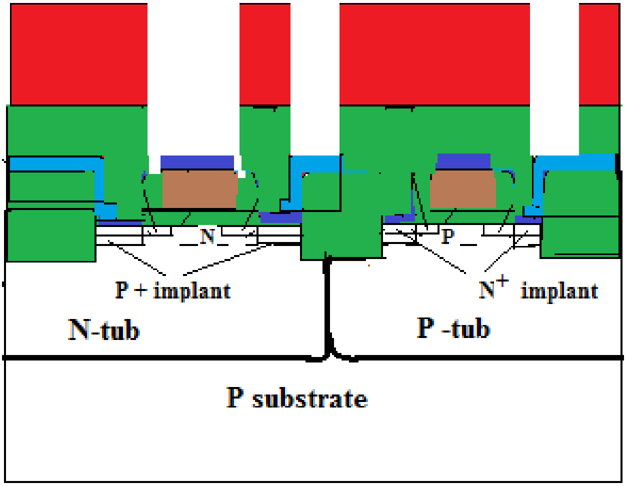


Figure 6.78. KPR is spun over the entire wafer. Mask 12 is used to make vias through which contacts will be established between Metal Level 1 and contact pads of the devices. The deposited SiO2 layer is etched out in vias region to allow connections to the silicon, polysilicon and to local interconnects.

After making the windows selectively through Mask 12, we etch SiO_2 in a plasma. After selectively etching the oxide in the vias, the fixed KPR is stripped off by oxygen plasma etching. We get the structure as shown in Figure 6.79.

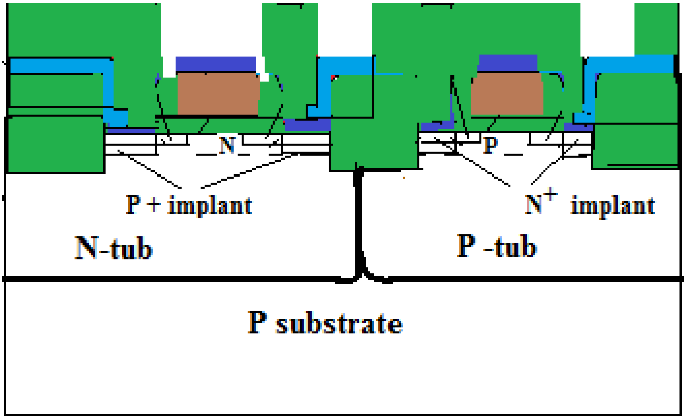


Figure 6.79. After the SiO2 is etched off the vias or contact holes, fixed KPR is stripped off.

We wish to maintain the planar surface as we add Metal Level 1 and Metal Level 2. There are many process flows for doing this but the commonest process flow is as follows:

A few tens of nm thick Ti/TiN bilayer is blanket deposited over the entire wafer area. This can be done by sputtering or by CVD. This bilayer provides a good adhesion to the SiO_2 and other underlying areas materials present in the structure at this point. The TiN acts as a barrier layer between upper Metal Level 1 and lower Local Interconnect layers ewhich connect the active devices. After the bilayer deposition we do tungsten(W) blanket layer deposition by CVD. A typical reaction might be:

$WF_6 + 3H_2 \rightarrow W + 6HF$

The resulting structure is shown in Figure 6.80. We do CMP to remove W and Ti/TiN from all the portions except the vias as shown in Figure 6.81. The resulting plane surface will be the base over which the metal layer will be deposited and interconnection pattern of first Metal layer will be etched out.

Next Metal deposition is done on the entire wafer surface as shown in Figure 6.82.. The metal is Aluminum. Aluminum has a small percentage of Si and small percentage of Cu. Si is soluble in Al and if Si is not present in the Metal Layer it will be absorbed from underlying Si rich layers. Absorption may cause contact problems as well as reliability problems. Hence it is preferred to consciously add Si to Metal Layer 1. Cu is added to prevent electronmigration from Al thin films. Electronmigration will lead to break in the interconnection and hence create yield problem. Electron flow in te metal line , over time, can cause the metal atoms to migrate along crystal grain boundaries or along the metal/dielectric interfaces in quasi-random manner. Voids may develop in the metal lines as a result and cause the line resistance to increase or even become open circuited. So Copper has replaced Al as the interconnect material in advanced ICs.

Copper has excellent electronmigration reliability and 40% lower resistance than Al. Copper may be deposited by plating or CVD. Because dry etching of Cu is difficult, copper patterns are defined by a damascene process. This will be discussed in a later section.

Al Metal Layer is shown in light sky blue colour. By photolithography technique and Mask 13 the interconnection pattern in the Metal Level 1 is etched out as shown in Figure 6.83.

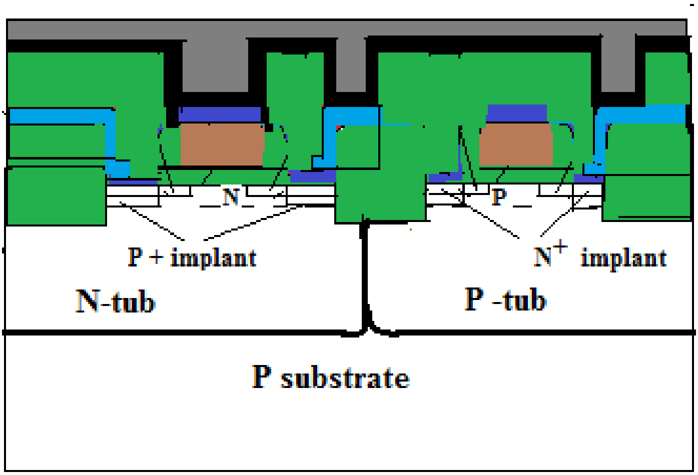


Figure 6.80. A thin Ti/TiN bilayer as a barrier and adhesion layer is deposited on the wafer by sputtering followed by deposition of W layer by CVD.

Ti/TiN bilayer is shown in black colour and W layer is shown as grey colour.

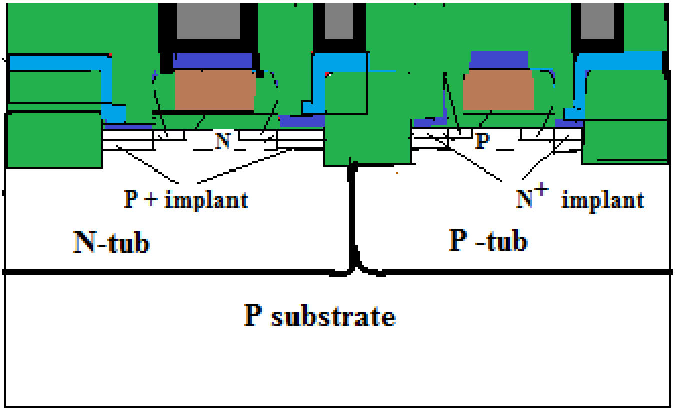


Figure 6.81. CMP is used to polish back the W and Ti/TiN bilayer leaving a planar surface on which the first Metal Level will be deposited.

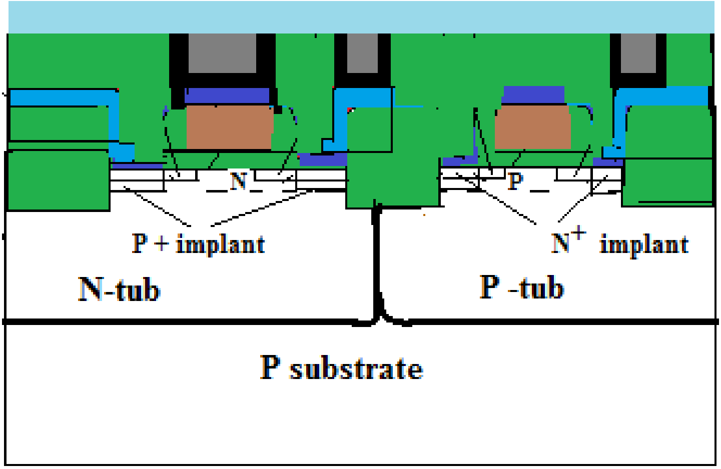


Figure 6.82. Metal Level 1 is deposited over the entire wafer area by sputtering. Metal Level 1 is Aluminum.

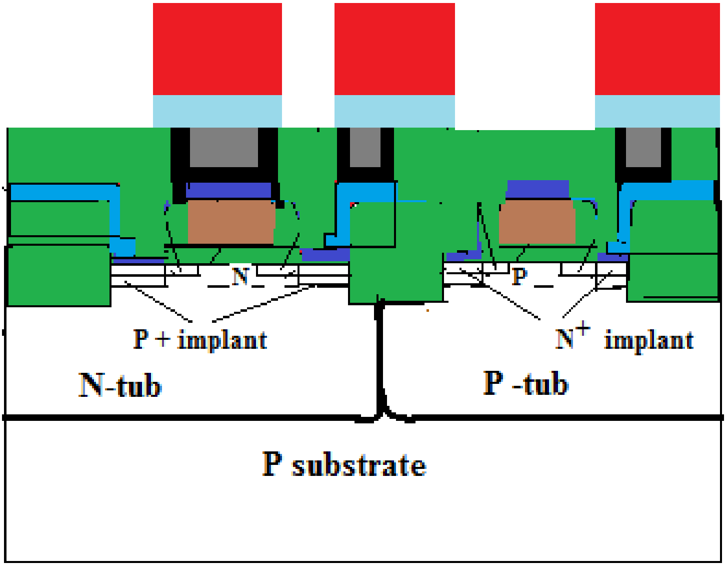


Figure 6.83. Aluminum is depositeed on the entire wafer by sputering. KPR is spun on the wafer and Mask 13 is used to define the first level of Metal interconnection pattern. The Al is then plasma etched

After the Metal Interconnect Pattern has been etched out, fixed KPR is stripped and again the entire wafer surface is covered with CVD Oxide layer. This Oxide layer is planarized, vias are defined by Mask 14 and vias are etched out. Now Metal Layer 2 is deposited again by sputtering and interconnect pattern is defined using Mask 15 and the redundant portion of the Metal is etched out by plasma etching. The final interconnect pattern is covered by Pasivation layer either of SiO_2 or Si_3N_4. This top layer is used to protect the chip during mechanical handling while packaging. This passivation layer will also protect against ambient contamination (Na^+ and K^+).

After the final processing step, the wafer is subjected to a low level thermal cycle of 400 to 450°C for about 30 minutes in forming gas (10% H_2 in N_2). This helps alloy the metal contacts to the contact pads through vias. This also helps reduce the electrical charges associated with Si/SiO_2 interface. The resulting structure is shown in Figure 6.84.

“To bulid complex and dense circuits , **the multi-level metallization** structure shown in Figure 6.84 is routinely employed. Up to about 10 metal layers may be used. The metal thickness may range from a small fraction of a micron to several; microns. The thinner interconnects route signals while the thicker layers serve as power lines.”

“The dielectric material between the metal layers used to be SiO_2. It has been supplemented with **low-k dielectrics** which often contain carbon or fluorine, and are designed to have much lower dielectric constants(k) than SiO_2. Lower k leads to lower capacitances between interconnects. This is highly desirable because capacitance in a circuit slows down the circuit speed, raises power consumption and introduces **cross talk** between neighboring interconnect lines.

Since a large number of metal layers and process steps are involved, making the interconnects *consumes a large part of the IC fabrication budget*. This part of the fabrication process is called the **back-end process**. In contrast , the steps used to produce the transistors are called the **front-end process**.”

[The quotes have been taken from the following book:

“Modern Semiconductor Devices for Integrated Circuits” by Chenming Calvin Hu, Pearson 2010]

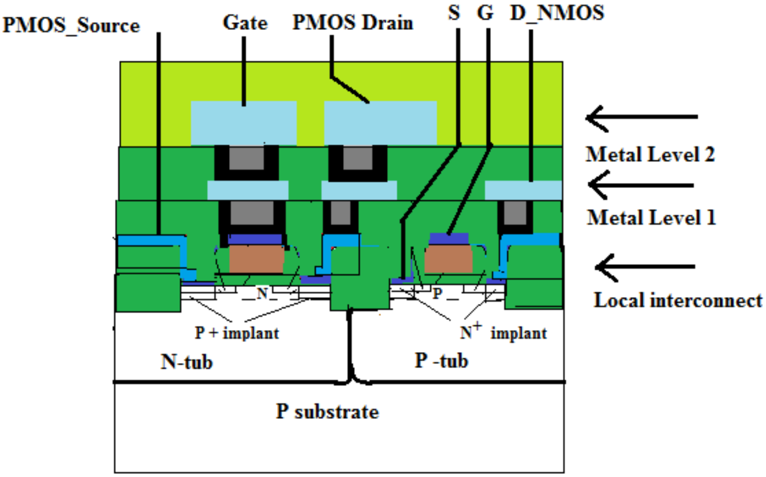


Figure 6.84. The steps to form the second level of Al interconnect follow the same route as we did for Metal Level 1. Mask 14 is used to define the via holes between Metal 1 and Metal 2. Mask 15 is used to define the interconnect pattern in Metal Level 2. The last step is deposition of a final pasivation layer of Si3N4 by PECVD. The last Mask 16 is used to open holes in this pasivation layer for access to banding pads of the circuit.

Class Test 1_25th_March_2011
Apollo Guidance Computer used RTL_3-input-wide-NOR-Gate in Integrated Circuit form. In Class Test 1, students have been asked to design the layout of the circuit.

Class Test 1_25.02.2011

Give the design, layout and fabrication steps for the dual-3-INPUT wide RTL-NOR Gate as shown in the Figure 1. This IC Chip was used successfully in Apollo Guidance Computer.

The **Apollo Guidance Computer** (**AGC**) provided onboard computation and control for guidance, navigation, and control of the Command Module (CM) and Lunar Module (LM) spacecraft of the Apollo program. It is notable for having been one of the first IC-based computers.

Each flight to the Moon (with the exception of Apollo 8, which didn't take a Lunar Module on its lunar orbit mission) had two AGCs, one each in the Command Module and the Lunar Module. The AGC in the Command Module was at the center of that spacecraft's guidance & navigation system (G&C). The AGC in the Lunar Module ran its Primary Guidance, Navigation and Control System, called by the acronym **PGNCS** (pronounced *pings*).

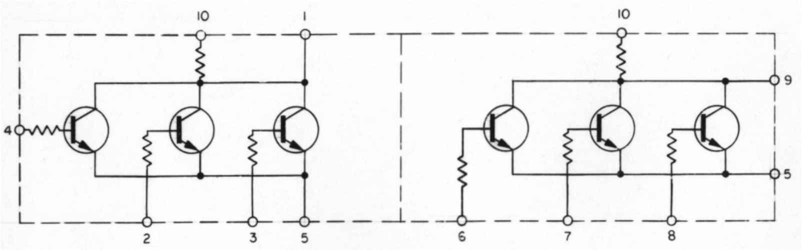
The AGC was designed at the MIT Instrumentation

Laboratory under Charles Stark Draper, with hardware design led by Eldon C. Hall.[1] Early architectural work came from J.H. Laning Jr., Albert Hopkins, Ramon Alonso,[2][3] and Hugh Blair-Smith.[4] The flight hardware was fabricated by Raytheon, whose Herb Thaler[5] was also on the architectural team.

The Apollo flight computer was the first to use integrated circuits (ICs). While the Block I version used 4,100 ICs, each containing a single 3-input NOR gate, the later Block II version (used in the crewed flights) used 2,800 ICs, each with two 3-input NOR gates.[1]:34 The ICs, from Fairchild Semiconductor, were implemented using resistor-transistor logic (RTL) in a flat-pack. They were connected via wire wrap, and the wiring was then embedded in cast epoxy plastic. The use of a single type of IC (the dual NOR3) throughout the AGC avoided problems that plagued another early IC computer design, the Minuteman IIguidance computer, which used a mix of diode-transistor logic and diode logic gates.

PGNC generated unanticipated warnings during Apollo 11's lunar descent, with the AGC showing a *1201 alarm* ("Executive overflow - no vacant areas") and a *1202 alarm* ("Executive overflow - no core sets").[8] The cause was a rapid, steady stream of spurious cycle steals from the rendezvous radar, intentionally left on standby during the descent in case it was needed for an abort. [9][10]

During this part of the approach the processor would normally be almost 85% loaded. The extra 6400 cycle steals per second added the equivalent of 13% load, leaving just enough time for all scheduled tasks to run to completion. Five minutes into the descent Buzz Aldrin gave the computer the command 1668 which instructed it to calculate and display DELTAH (the difference between altitude sensed by the radar and the computed altitude). This added an additional 10% to the processor workload causing executive overflow and a 1202 alarm. After being given the "GO" from Houston Aldrin entered 1668 again and another 1202 alarm occurred. When reporting the second alarm Aldrin added the comment "It appears to come up when we have a 1668 up". Happily for Apollo 11 the AGC software had been designed with priority scheduling. Just as it had been designed to do, the software automatically recovered, deleting lower priority tasks including the 1668 display task, to complete its critical guidance and control tasks. Guidance controller Steve Bales and his support team that included Jack Garman issued several "GO" calls and the landing was successful. For his role, Bales received the US Medal of Freedom on behalf of the entire control center team and the three Apollo astronauts.[11]



SCHEMATIC

THIS SCHEMATIC IS REPRESENTATIVE OF THE ELECTRICAL CHARACTERISTICS ONLY. THE PHYSICAL CIRCUITRY IS ENTIRELY CONTAINED WITHIN A MICRO NOR GATE FLAT PACK

Figure 1.

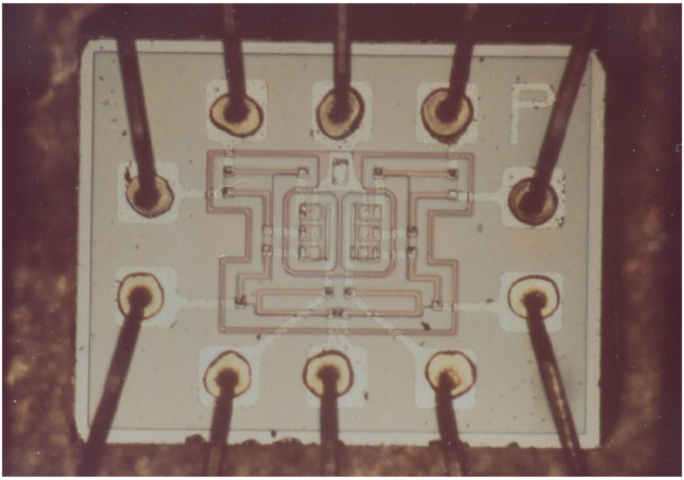


Figure 2. Monolithic Circuit Dual 3-input-wide_RTL-NOR-Gate interconnections and bonding pad layout.

Figure 2 is the photomicrograph of the layout.

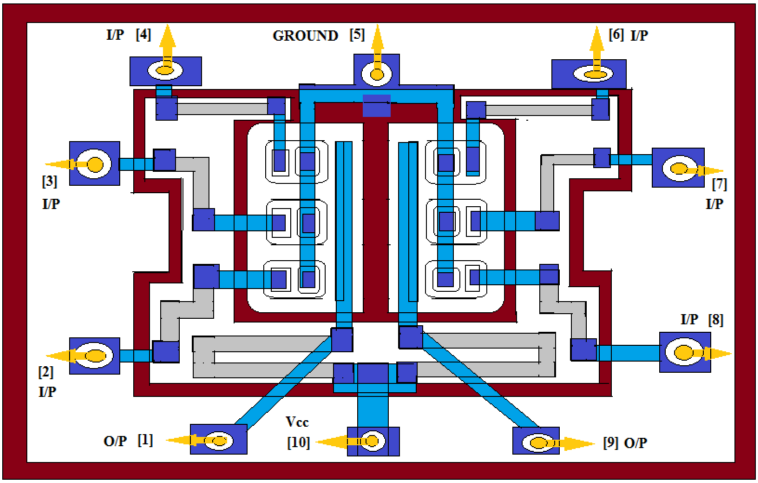


Figure 3.This is manually drawn several times magnified layout of the monolithic circuit.

- Deep Brown is the isolation boundary.
- Navy Blue is the bonding pad.
- Sky Blue is the metallization.
- Golden yellow is the gold wire which is bonded to the peripheral bonding pad on one end and to the terminal of the IC package at the other end. This bonding is done either by Thermo-compression technique or by Ultra-sonic bonding technique.
- The die or the individual chip is mounted on ceramic substrate by means of solder glass. The ceramic substrate is attached to a suitable header by means of a high temperature eutectic solder or the die can be mounted in all ceramic flat package.
- The header is hermetically sealed by a suitable

packaging material.

Design rules for Monolithic Circuit Layout.

1. Redraw the circuit to satisfy the required pin connections with minimum crossovers.
2. Determine the number of isolation islands from collector potential consideration and reduce the area as much as you can.
3. Place all the resistances in the same isolation island and this isolated island must be returned to the most positive voltage.
4. Resistances are realized by an elongated strip of P diffusion in N-epi layer. In order that resistances may be isolated from N-epi layer and hence from one another, N-epi layer corresponding to the island containing the resistances must be at the most positive potential namely Vcc.
5. The substrate must be at the most negative potential. In this case the GROUND potential.
6. Allow the isolation border to be equal to twice the epilayer thickness.
7. Use 1-mil width for diffused N+ emitter region.
8. Use ½ mil width for base contact and spacing.
9. Use ½ mil width for collector contact and spacing.
10. For resistances use widest possible design.
11. Keep metallization short and wide.

Solution of the problem:

1. Number of islands: Every distinct collector voltage will have one isolated island assigned to it. Here we have three transistors in one gate but at the same collector voltage therefore the two gates wil have two islands containing the two sets of three transistors.
2. This dual gate will have a third island which will be at the highest Positive Voltage Vcc and will contain the two sets of four resistances. Pin number 10 is connected to Vcc and is connected to the N-epilayer island containing all the resistances
3. The P_- substrate will be held at the most negative voltage which in this case is Ground. Pin number 5 is the ground terminal and is connected to a bonding pad connected to the isolation boundary(deep brown colour) as shown.
4. The N-epilayer containing the transistors has one elongated N+ diffusion to which Aluminum contact pad is made. This is the collector Contact Pad. This collector Contact Pad is connected to the Output Pad as well as to one end of the resistance R_C .
5. Similarly emitter contact pads are connected together and connected to ground.
6. The base Contact Pads are connected to R_B's which in turn are connected to the three input pads.

7. Fabrication Steps are:

1. Oxidation of the substrate of resistivity 10 Ω-cm.
2. Using Mask 1 , buried layer diffusion of Arsenic is done with sheet resistivity of 1 Ω/sq.
3. Oxide layer is completely etched away and epitaxial layer of N type Silicon is grown of resistivity 0.1Ω-cm.
4. Reoxidize and using Mask 2, isolation diffusion P-type is done to create isolated N-epi islands. Each island will contain BJTs or Diodes with collectors at one voltage.
5. Reoxidize and open windows by Mask 3 for Boron base diffusion and for Resistance diffusion ($3R_{BS}$ and R_{C}). All the resistances are located on the island which has its N-epi layer at V_{CC}.
6. Reoxidize and open windows by Mask 4 for Phosphorous emitter diffusion and for collector contact diffusion. Three separate emitters have to be defined and one elongated collector contact is defined.
7. Reoxidize and open apertures by Mask 5 for contact pads to 3 emitters, 3 bases and one elongated collector per gate. The contact pads are made to the two ends of the four resistances also per gate.
8. Final metallization is done covering the whole wafer and using Mask 6, the interconnection pattern is etched out as shown in Figure 3.

9. Golden yellow in Figgure 3 is the gold wire which is bonded to the peripheral bonding pad on one end and to the terminal of the IC package at the other end. This bonding is done either by Thermo-compression technique or by Ultra-sonic bonding technique.
10. The die or the individual chip is mounted on ceramic substrate by means of solder glass. The ceramic substrate is attached to a suitable header by means of a high temperature eutectic solder or the die can be mounted in all ceramic flat package.
11. The header is hermetically sealed by a suitable packaging material.

SSPD_Chapter 6_Part 7_Device Simulation1
SSPD_Chapter 6_Part 7_Device Simulation 1 introduces to the use of ATLAS softeware and diode simulation is carried out and its Avalanche Breakdown is analyzed

SSPD_Chapter 6_Part 7_Device Simulation1.

(Atlas User Manual has been used for preparing this Section. support@silvaco.com may be contacted for any assistance required while using their software.)

In this Section Part 7 we will use Device Simulation Software for studying the device testing and characterization in a virtual lab. In this particular instance we will use ATLAS for Device Simulation and Characterization.

The information presented is based on the assumptions that the reader is (1) familiar with the basic terminology of semiconductor processing and semiconductor device operation, and (2) understands the basic operation of the computer hardware and operating system being employed.

ATLAS is a modular and extensible framework for one, two and three dimensional semiconductor device simulation. It is implemented using modern software engineering practices that promote reliability, maintainability, and extensibility.

Products that use the ATLAS Framework meet the device simulation needs of all semiconductor application areas.

ATLAS provides general capabilities for physically-based two (2D) and three-dimensional (3D) simulation of semiconductor devices. ATLAS is designed to be used in conjunction with the VWF (Virtual Wafer Fab) INTERACTIVE TOOLS. The VWF INTERACTIVE TOOLS are DECKBUILD, TONYPLOT, DEVEDIT, MASKVIEWS, and OPTIMIZER.

ATLAS is supplied with numerous examples that can be accessed through DECKBUILD. These examples demonstrate most of ATLAS's capabilities. The input files that are provided with the examples given in Part 7.1 are an excellent starting point for developing your own input files. Part 7.1 gives "Getting Started with ATLAS" which contains a section "Accessing The Examples".

7.1. Features and capabilities of ATLAS.

ATLAS provides a comprehensive set of physical models, including:

- DC, AC small-signal, and full time-dependency.

- Drift-diffusion transport models.

- Energy balance and Hydrodynamic transport models.

- Lattice heating and heatsinks.
- Graded and abrupt heterojunctions.
- Optoelectronic interactions with general ray tracing.
- Amorphous and polycrystalline materials.
- General circuit environments.
- Stimulated emission and radiation
- Fermi-Dirac and Boltzmann statistics.
- Advanced mobility models.
- Heavy doping effects.
- Full acceptor and donor trap dynamics
- Ohmic, Schottky, and insulating contacts.
- SRH, radiative, Auger, and surface recombination.
- Impact ionization (local and non-local).
- Floating gates.
- Band-to-band and Fowler-Nordheim tunneling.
- Hot carrier injection.

- Quantum transport models

- Thermionic emission currents.

ATLAS uses powerful numerical techniques, including:

- Accurate and robust discretization techniques.

- Gummel, Newton, and block-Newton nonlinear iteration strategies.

- Efficient solvers, both direct and iterative, for linear subproblems.

- Powerful initial guess strategies.

- Small-signal calculation techniques that converge at all frequencies.

- Stable and accurate time integration.

ATLAS is often used in conjunction with the ATHENA process simulator. ATHENA predicts the physical structures that result from processing steps. The resulting physical structures are used as input by ATLAS, which then predicts the electrical characteristics associated with specified bias conditions. The combination of ATHENA and ATLAS makes it possible to determine the impact of process parameters on device characteristics.

The electrical characteristics predicted by ATLAS can be used as input by the UTMOST device characterization and SPICE modeling software. Compact models based on simulated device characteristics can then be supplied to circuit designers for preliminary circuit design. Combining ATHENA, ATLAS, UTMOST, and SMARTSPICE makes it possible to predict the impact of process parameters on circuit characteristics.

7.2. Physically based device simulator.

ATLAS is a physically-based device simulator. Physically-based device simulation is not a familiar concept for all engineers. This section will briefly describe this type of simulation.

Physically-based device simulators predict the electrical characteristics that are associated with specified physical structures and bias conditions. This is achieved by approximating the operation of a device onto a two or three dimensional grid, consisting of a number of grid points called nodes. By applying a set of differential equations, derived from Maxwells laws, onto this grid you can simulate the transport of carriers through a structure. This means that the electrical performance of a device can now be modelled in DC, AC or transient modes of operation.

The Maxwell Laws are:

1. Divergence of electric flux from the control volume = Charge Density in the control volume (Poisson's Equation) or Divergence of electric flux = zero (Laplace's Equation);
2. Divergence of electrons from a control volume = difference in generation rate and recombination rate of electrons in the control volume;
3. Divergence of holes from a control volume = difference in generation rate and recombination rate of holes in the control volume;
4. Total electron current density = drift electron flux density + diffusion electron flux density;
5. Total hole current density = drift hole flux density + diffusion hole flux density;

First three are coupled , non-linear second order partial differential equations and last two

are transport equations.

Solution of these five equations give the simulated predictions.

The intersection of the grid lines give the nodes. At node (i,j) there are three unknowns v_{ij} , n_{ij} and p_{ij} .Initial intelligent guess is made of v_{ij} , n_{ij} and p_{ij} . Using these values we solve the Poisson equations and the continuity equations for electron and holes until we get converging results.

There are two methods for solving for these three unknowns: Coupled Method and Uncoupled Method or Sequential Method.

Table 7.1. Comparative study of the Coupled and Uncoupled Methods.

Coupled Method	Uncoupled (sequential) method
Generalization of Newton method	Basis of Gummel’s method
Large memory required	Memory efficient
Converges rapidly	May converge rapidly at low DC bias
Initial guess is critical	Converges slowly at large DC bias

There are three physically-based simulation. These are:

- It is predictive.
- It provides insight.
- It conveniently captures and visualizes theoretical knowledge.

Physically-based simulation is different from empirical modelling. The goal of empirical modelling is to obtain analytic formulae that approximate existing data with good accuracy and minimum complexity. Empirical models provide efficient approximation and interpolation. They do not provide insight, or predictive capabilities, or encapsulation of theoretical knowledge.

Physically-based simulation has become very important for two reasons. One, it is almost always much quicker and cheaper than performing experiments. Two, it provides information that is difficult or impossible to measure.

The drawbacks of physically-based simulation are that all the relevant physics must be incorporated into a simulator, and numerical procedures must be implemented to solve the associated equations. These tasks have been taken care of for users of ATLAS.

Those who use physically-based device simulation tools must specify the problem to be simulated. In ATLAS, specify device simulation problems by defining:

- The physical structure to be simulated.

- The physical models to be used.

- The bias conditions for which electrical characteristics are to be simulated.

7.3. Application of ATLAS.

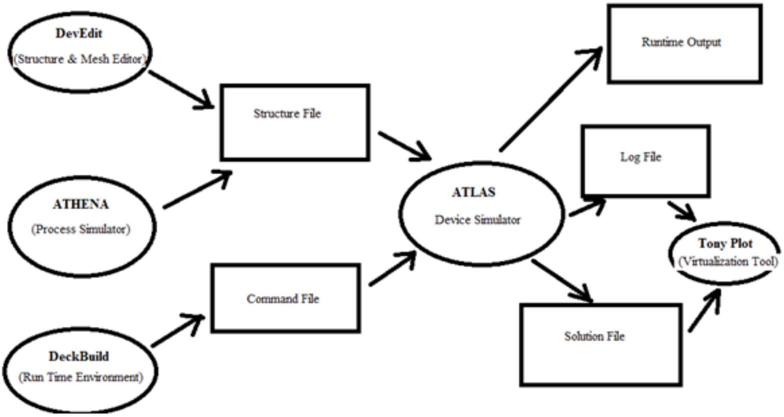


Figure 7.1. ATLAS INPUTS AND OUTPUTS.

Figure 7.1 shows the types of information that flow in and out of ATLAS. Most ATLAS simulations use two input files. The first input file is a text file that contains commands for ATLAS to execute. The second input file is a structure file that defines the structure that will be simulated.

ATLAS produces three types of output files. The first type of output file is the run-time output, which gives you the progress and the error and warning messages as the simulation proceeds. The second type of output file is the log file, which stores all terminal voltages and currents from the device analysis. The third type of output file is the solution file, which stores 2D and 3D data relating to the values of solution variables within the device at a given bias point.

7.3.1: Modes of Operation

ATLAS is normally used in conjunction with the DECKBUILD run-time environment, which supports both interactive and batch mode operation. We strongly recommend that you always run ATLAS within DECKBUILD. In this section, we present the basic information you need to run ATLAS in DECKBUILD.

7.3.1.1: Interactive Mode With DeckBuild

To start ATLAS in DECKBUILD, type:

deckbuild - as

at the UNIX system command prompt. The command line option, *-as*, instructs DECKBUILD to start ATLAS as the default simulator.

If you want to start from an existing input file, start DECKBUILD by typing:

deckbuild - as <input filename>

The run-time output shows the execution of each ATLAS command and includes error messages, warnings, extracted parameters, and other important output for evaluating each ATLAS run. When ATLAS runs in this mode, the run-time output is sent to the output section of the DeckBuild Window and can be saved as needed. Therefore, you don’t need to save the run-time output explicitly. The following command line, however, specifies the name of a file that will be used for storing the run-time output.

```
deckbuild -as <input filename> - outfile <output filename>
```

In this case, the run-time output is sent to the output file and to the output section of the DeckBuild Window.

7.3.1.2: Running ATLAS inside Deckbuild

Each ATLAS run inside DECKBUILD should start with the line:

```
go atlas
```

A single input file may contain several ATLAS runs each separated with a go atlas line. Input files within DECKBUILD may also contain runs from other programs such as ATHENA or DEVEDIT along with the ATLAS runs.

Running a given version number of ATLAS

The go statement can be modified to provide parameters for the ATLAS run. To run version 4.3.0.R, the syntax is:

```
go atlas simflags = ”-V 4.3.0.R”
```

Starting Parallel ATLAS

The -P option is used to set the number of processors to use in a parallel ATLAS run. If the number set by -P is greater than the number of processors available or than the number of parallel thread licenses, the number is automatically reduced to this cap number. To run on 4 processors, use:

```
go atlas simflags=”-V 4.3.2.C -P 4”
```

Note: The standard examples supplied with ATLAS will not run correctly outside of DECKBUILD.

7.3.1.3: Accessing The Examples

ATLAS has more than 300 standard examples that demonstrate how the program is used to simulate different technologies. These examples are a good starting point for creating your own simulations. The examples are accessed from the menu system in DECKBUILD. To select and load an example:

1. Start DECKBUILD with ATLAS as the simulator, which is described in the previous section.

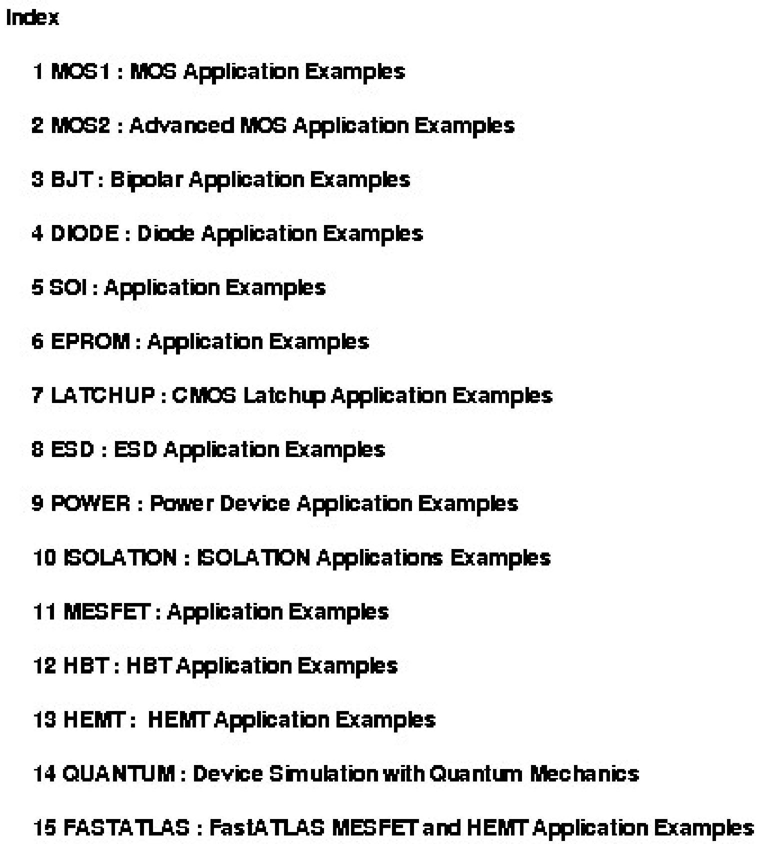
2. Use left mouse button to pull down the Main Control menu.

3. Select Examples. An index will then appear in a Deckbuild Examples Window (see Figure 7.2).

The examples are divided by technology or technology group. For instance, the most common technologies are individually listed (e.g., MOS, BJT), while others are grouped with similar devices (e.g., IGBT and LDMOS are under POWER, and solar cell and photodiode are under OPTOELECTRONICS).

4. Choose the technology by double clicking the left mouse button over that item. A list of examples for that technology will appear. These examples typically illustrate different devices, applications, or types of simulation.

You can also search for an example by selecting the Index button. Wildcards can be used in the search.



Index

1 MOS1 : MOS Application Examples

2 MOS2 : Advanced MOS Application Examples

3 BJT : Bipolar Application Examples

4 DIODE : Diode Application Examples

5 SOI : Application Examples

6 EPROM : Application Examples

7 LATCHUP : CMOS Latchup Application Examples

8 ESD : ESD Application Examples

9 POWER : Power Device Application Examples

10 ISOLATION : ISOLATION Applications Examples

11 MESFET : Application Examples

12 HBT : HBT Application Examples

13 HEMT : HEMT Application Examples

14 QUANTUM : Device Simulation with Quantum Mechanics

15 FASTATLAS : FastATLAS MESFET and HEMT Application Examples

Figure 7.2 Examples Index in DeckBuild

5. Choose a particular example by double clicking the left mouse button over that item in the list. A text description of the example will appear in the window. This text describes the important physical mechanisms in the simulation, and details of the ATLAS syntax used. You should read this information before proceeding.

6. Press the Load Example button. The input

command file for the example will be copied into your current working directory together with any associated files. A copy of the command file will be loaded into DECKBUILD. Note that the Load Example button remains faded out until this step is performed correctly.

7. Press the Run button in the middle frame of the DECKBUILD application window to run the example. Alternatively, most examples are supplied with results that are copied into the current working directory along with the input file. To view the results, select (highlight) the name of the results file and select Tools-Plot. Details on using TONYPLOT can be found in the VWF INTERACTIVE TOOLS USER'S MANUALS.

```
go atlas

TITLE PN Diode Breakdown Simulation with curve tracing algorithm

# SILVACO International 1996

mesh

x.m l=0.0 spac=1.0#structure

x.m l=1.0 spac=1.0#structure

y.m l=0 spac=1.0#structure
```

```
y.m l=5.0 spac=0.005#structure

y.m l=15 spac=2#structure

region num=1 silicon#material

electrode top name=emitter#structure

electrode bottom name=base#structure

doping uniform conc=5e17 p.type #doping

doping uniform n.type conc=1.e20 x.l=0. x.r=1
y.t=0.0 y.b=5.0# doping

save outf=diodeex03_0.str # solution specification

#tonyplot diodeex03_0.str -set diodeex03_0.set#
result analysis

models srh conmob bgn auger fldmob # models

impact crowell # models

solve init # solution specification

solve # solution specification

solve vemitter=0.1 # solution specification

method newton climit=1.e-4 #numerical method
selection
```

```
curvetrace end.val=1e-4 contr.name=emitter curr.cont mincur=1e-13 nextst.ratio=1.2

# result analysis

log outf=diodeex03.log # result analysis

solve curvetrace # result analysis

tonyplot diodeex03.log -set diodeex03.set # result analysis

quit
```

The first part of the program (Deckbuild up) gives the 2-D mesh description of the cross-sectional view of the emitter-base diode. The upper part of the cross-sectional view is Emitter (N type) which has an uniform doping of 10^{20} Phosphorous atoms per c.c. and lower part is Base (P-type) which has an uniform doping of 5×10^{17}Boron atoms per c.c. as shown in Figure 7.3.

It uses the models given in Table 7.2.

In Method statement we give the Numerical Method adopted for carrying the calculations.

In weakly coupled systems we use Gummel Numerical Method where we regard the system as decoupled and we analyze the nodes one by one. This has linear convergence.

In strongly coupled systems we use Newton Numerical Method . This has quadratic convergence.

In Block system we start as decoupled system and then we use coupling.

In the example above we have used Newton Method.

We have carried out two simulations.

Table 7.2. The Models used in this program

Model	Syntex	Notes
Shockley-Hall-Reed indirect recombination	SRH	This describes the method of indirect recombination mediated by traps and defects.
Concentration Dependent Mobility	CONMOB	Impurity scattering is considered along with lattice scattering.
Band Gap Narrowing	bgn	Emitter has heavy doping

		leading to band gap narrowing this adversely effects the emitter injection efficiency.(This has been elaborately explained in Appendix IV of IC journey from Micro to Nano Era. This Appendix is a part of SSPD collection)
Auger Direct Recombination	auger	Whenever carrier concentrations are high they directly recombine. In that case trap mediation is not required.
Field Dependent Mobility	fldmob	At high fields where drift velocity is comparable to the thermal velocity the drift velocity saturates

		to scatter limited velocity.
Crowell-Sze model	impact crowell	Impact ionization model has to be used while modelling the Avalanche Breakdown

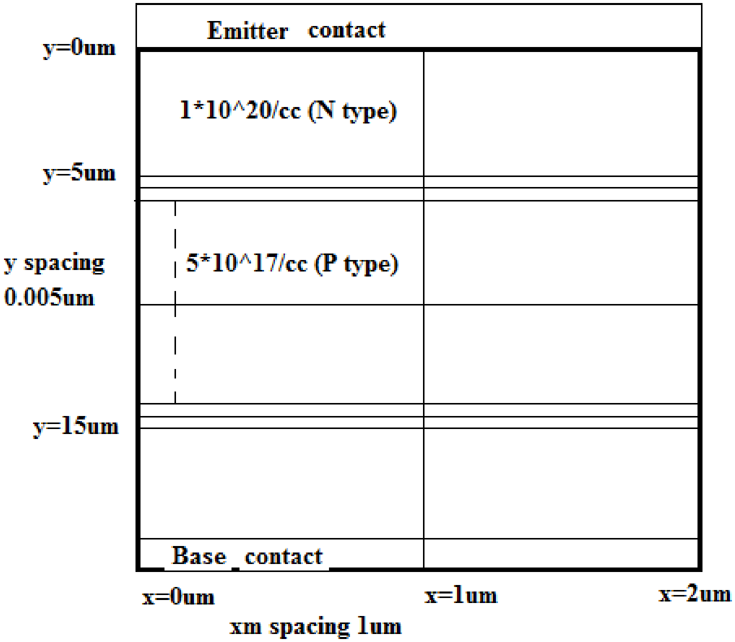


Figure 2. Device 2-D cross-sectional structure described by X-Y grid lines. The location and spacing of the grid lines are shown. The critical part or the active part of the device cross-section are covered by fine mesh. Remaining part is covered by coarse mesh.

Figure 7.3. 2D Mesh describing the cross-sectional area of the Device.

It should be noted that lightly doped portion base is being covered by fine mesh whereas heavily doped emitter is covered by coarse mesh. In One-sided step Junction (N+P) minority carriers electron in P-side carry the major part of the diode current hence minority carrier distribution in P-side have to be more accurately analyzed hence base is being covered by a finer mesh.

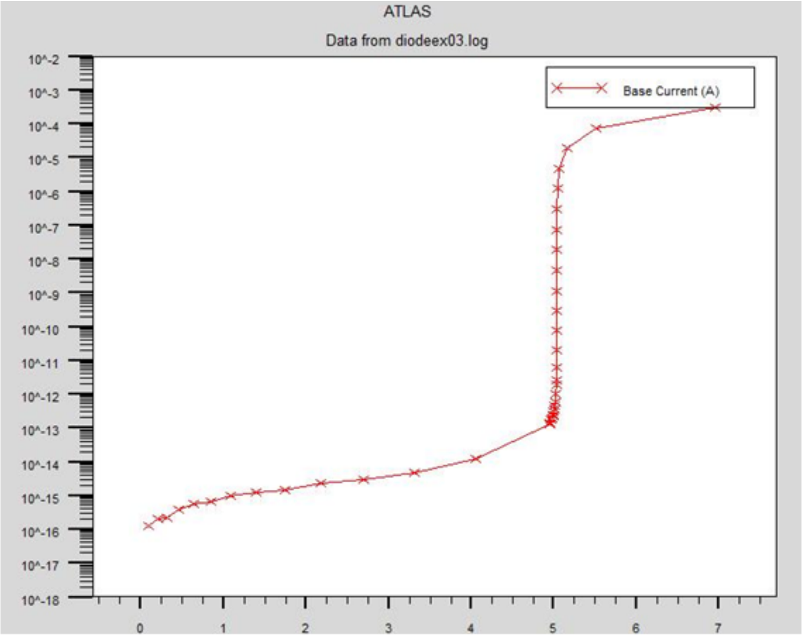


Figure 3. Simulation Number 1 tony plot. BVz = 5V

go atlas

TITLE PN Diode Breakdown Simulation with curve tracing algorithm

```
# SILVACO International 1996

mesh

x.m l=0.0 spac=1.0

x.m l=1.0 spac=1.0

y.m l=0 spac=1.0

y.m l=5.0 spac=0.005

y.m l=15 spac=2

region num=1 silicon

electrode top name=emitter

electrode bottom name=base

doping uniform conc=5e16 p.type

doping uniform n.type conc=1.e20 x.l=0. x.r=1 y.t=0.0 y.b=5.0

save outf=diodeex03_0.str

#tonyplot diodeex03_0.str -set diodeex03_0.set

models srh conmob bgn auger fldmob

impact crowell
```

```
solve init

solve

solve vemitter=0.1

method newton climit=1.e-4

curvetrace end.val=1e-4 contr.name=emitter
curr.cont mincur=1e-13 nextst.ratio=1.2

log outf=diodeex03.log

solve curvetrace

tonyplot diodeex03.log -set diodeex03.set

quit
```

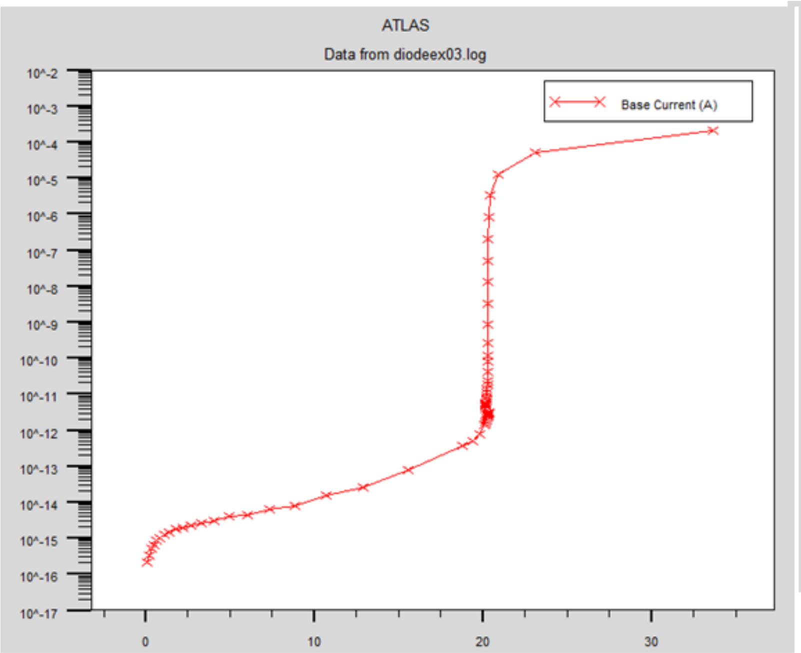


Figure 4. Simulation Number 2 tony plot. BVa = 20V.

7.3.1.4. Discussion of the two simulations carried out in the previous section.

We have just carried out two device simulations for PN Junction breakdown with curve tracing algorithm.

In first device, base had a concentration of 5×10^{17}Boron atoms per cc. This corresponded to a reverse Avalanche Breakdown of 5V. At an emitter doping of 1×10^{20}/cc we should have invoked 'Tunnelling Model' and obtained Zener Break down.

In second device base had a concentration of 5×10^{16} Boron atoms per cc. This corresponded to a reverse Avalanche Breakdown of 20V.

SSPD_Chapter 6_Part 7_Device Simulation2
SSPD_Chapter 6_Part 7_Device Simulation 2 deal with the program for BJT simulation and generation of Ic-Vce family of curves. It gives the programming description of the structure and doping of Vertical NPN transistor.

SSPD_Chapter 6_Part 7_ Device Simulation2_to be continued

7.4. Structure and Electrical Characterization of NPN Transistor.

The input file contains a sequence of statements. Each statement consists of a keyword that identifies the statement and a set of parameters. The general format is:

<STATEMENT> <PARAMETER>=<VALUE>

With a few exceptions the input syntax is not case sensitive. One important exception is that commands described in this manual as being executed by DECKBUILD rather than ATLAS are case sensitive. These include: EXTRACT, SET, GO, and SYSTEM. Also, filenames for input and output under UNIX are case sensitive.

For any <STATEMENT>, ATLAS may have our different types for the <VALUE> parameter. These are: Real, Integer, Character, and Logical.

An example of a statement line is:

```
DOPING UNIFORM N.TYPE
CONCENTRATION=1.0e16 REGION=1
OUTFILE=my.dop
```

The statement is DOPING. All other items are parameters of the DOPING statement. UNIFORM and N.TYPE are logical parameters. Their presence on the line sets their values to true, otherwise they take their default values (usually false). CONCENTRATION is a Real parameter and takes floating point numbers as input values. REGION is an Integer parameter taking only integer numbers as input. OUTFILE is a Character parameter type taking strings as input.

The statement keyword must come first but after this, the order of parameters within a statement is not important.

You only need to use enough letters of any parameter to distinguish it from any other parameter on the same statement. Thus, CONCENTRATION can be shortened to CONC. REGION, but can't be shortened to R, since there's a parameter called RATIO associated with the DOPING statement.

Logical parameters can be explicitly set to false by preceding them with the ˆ symbol. Any line beginning with a # is ignored. These lines are used as comments.

ATLAS can read up to 256 characters on one line. But it is best to spread long input statements over several lines to make the input file more readable. The \ character at the end of a line indicates continuation.

7.4.1: The Order of ATLAS Commands

The order in which statements occur in an ATLAS input file is important. There are five groups of statements that must occur in the correct order (see Table 7.2). Failure to do so usually causes an error message to appear, which could lead to incorrect operation or termination of the program. For example, if the material parameters or models are set in the wrong order, then they may not be used in the calculations.

The order of statements within the mesh definition, structural definition, and solution groups is also important. Failure to place these statements in proper order may also cause incorrect operation or termination of the program.

Table 7.2. ATLAS Command Group with Primary Statements in each Group.

No.	Group	Statements
1	Structure Specification	MESH, REGION, ELECTRODE, DOPING
2	Material Models Specification	MATERIAL MODEL CONTACT INTERFACE
3	Numerical Method Selection	METHOD
4	Solution Specification	LOG SOLVE LOAD SAVE
5	Result Analysis	EXTRACT TONY PLOT

7.4.2.: Interface FromDevEdit

A 2D or 3D structure created by DEVEDIT (device edit) can be read into ATLAS using the following statement.

MESH INF=<structure filename>

This statement loads in the mesh, geometry, electrode positions, and doping of the structure. ATLAS will automatically determine whether the mesh is 2D for S-PISCES or BLAZE, or 3D for DEVICE3D or BLAZE3D.

If the structure coming from DEVEDIT were originally created by ATHENA, the electrodes should be defined in ATHENA as described in the previous section. If the structure is created in DEVEDIT, the electrode regions should be defined in the Region/Add menu of DEVEDIT.

7.4.3.: Using The Command Language To Define A Structure

To define a device through the ATLAS command language, you must first define a mesh. This mesh or grid covers the physical simulation domain. The mesh is defined by a series of horizontal and vertical lines and the spacing between them. Then, regions within this mesh are allocated to different materials as required to construct the device. For example, the specification of a MOS device requires the specification of silicon and silicon dioxide regions. After the regions are defined, the location of electrodes is specified. The final step is to specify the doping in each region.

When using the command language to define a structure, the information described in the following four sub-sections must be specified in the order listed.

Specifying the Initial Mesh

The first statement must be:

```
MESH SPACE.MULT = <VALUE>
```

This is followed by a series of X.MESH and Y.MESH statements.

X.MESH LOCATION = <VALUE> SPACING = <VALUE>

.

Y.MESH LOCATION = <VALUE> SPACING = <VALUE>

.

The SPACE.MULT parameter value is used as a scaling factor for the mesh created by the X.MESH and Y.MESH statements. The default value is 1. Values greater than 1 will create a globally coarser mesh for fast simulation. Values less than 1 will create a globally finer mesh for increased accuracy. The X.MESH and Y.MESH statements are used to specify the locations in microns of vertical and horizontal lines, respectively, together with the vertical or horizontal spacing associated with that line. At least two mesh lines must be specified for each direction. ATLAS automatically inserts any new lines required to allow for gradual transitions in the spacing values between any adjacent lines. The X.MESH and Y.MESH statements must be listed in the order of increasing x and y. Both negative and positive values of x and y are allowed.

After an initial mesh has been defined, you can remove grid lines in specified regions. This is typically done in regions of the device where a coarse grid is expected to be sufficient such as the substrate. The removal of grid lines is accomplished using the ELIMINATE statement. The ELIMINATE statement removes every second mesh line in the specified direction from within a specified rectangle. For example, the statement:

```
ELIMINATE COLUMNS X.MIN=0 X.MAX=4
Y.MIN=0.0 Y.MAX=3
```

removes every second vertical grid line within the rectangle bounded by x=0, x=4, y=0 and y=3 microns.

7.4.4. Specifying Regions and Materials

Once the mesh is specified, every part of it must be assigned a material type. This is done with REGION statements. For example:

```
REGION number = <integer> <material_type>
<position parameters>
```

Region numbers must start at 1 and are increased for each subsequent region statement. You can have up to 200 different regions in ATLAS. A large number of materials is available. If a composition-dependent material type is defined, the x and y composition fractions can also be specified in the REGION statement.

The position parameters are specified in microns using the X.MIN, X.MAX, Y.MIN, and Y.MAX parameters. If the position parameters of a new statement overlap those of a previous REGION statement, the overlapped area is assigned as the material type of the new region. Make sure that materials are assigned to all mesh points in the structure. If this isn't done, error messages will appear and ATLAS won't run successfully.

You can use the MATERIAL statement to specify the material properties of the defined regions. But you must complete the entire mesh and doping definition before any MATERIAL statements can be used.

7.4.5. Specifying Electrodes

Once you have specified the regions and materials, define at least one electrode that contacts a semiconductor material. This is done with the ELECTRODE statement. For example:

ELECTRODE NAME=<electrode name> <position_ parameters>

You can specify up to 50 electrodes. The position parameters are specified in microns using the X.MIN, X.MAX, Y.MIN, and Y.MAX parameters. Multiple electrode statements may have the same electrode name. Nodes that are associated with the same electrode name are treated as being electrically connected.

Some shortcuts can be used when defining the location of an electrode. If no y coordinate parameters are specified, the electrode is assumed to be located on the top of the structure. You also can use the RIGHT, LEFT, TOP, and BOTTOM parameters to define the location. For example:

```
ELECTRODE NAME=SOURCE LEFT LENGTH=0.5
```

specifies the source electrode starts at the top left corner of the structure and extends to the right for the distance LENGTH.

7.4.6. Specifying Doping

You can specify analytical doping distributions, or have ATLAS read in profiles that come from either process simulation or experiment. You specify the doping using the DOPING statement. For example:

```
DOPING <distribution_type> <dopant_type>
<position_parameters>
```

Analytical Doping Profiles

Analytical doping profiles can have uniform or Gaussian forms. The parameters defining the analytical distribution are specified in the DOPING statement. Two examples are shown below with their combined effect shown in Figure 7.5.

```
DOPING UNIFORM CONCENTRATION=1E16
N.TYPE REGION=1

DOPING GAUSSIAN CONCENTRATION=1E18
CHARACTERISTIC=0.05 P.TYPE \ X.LEFT=0.0
X.RIGHT=1.0 PEAK=0.1
```

The first DOPING statement specifies a uniform n-type doping density of 10E16 cm-3 in the region that was previously labelled as region #1. The position parameters: X.MIN, X.MAX, Y.MIN, and Y.MAX can be used instead of a region number.

The second DOPING statement specifies a p-type Gaussian profile with a peak concentration of 10E18 cm-3. This statement specifies that the peak doping is located along a line from $x = 0$ to $x = 1$ microns. Perpendicular to the peak line, the doping drops off according to a Gaussian distribution with a standard deviation of 0.05 mm. At $x < 0$ or $x > 1$, the doping drops off laterally with a default standard deviation that is 70% of CHARACTERISTIC. This lateral roll-off can be altered with the RATIO.LATERAL parameter. If a Gaussian profile is being added to an area that was already defined with the opposite dopant type, you can use the JUNCTION parameter to specify the position of the junction depth instead of specifying the standard deviation using the CHARACTERISTIC parameter.

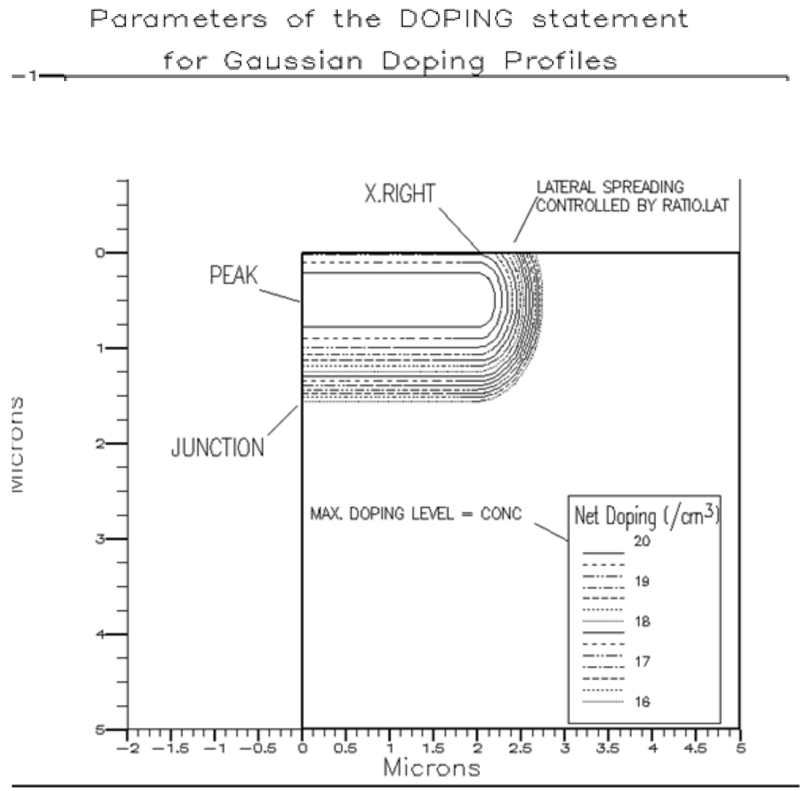


Figure 7.5. Parameters of the doping statement for Gaussian Doping Profiles.

7.4.7. Specifying Physical Models

Physical models are specified using the MODELS and IMPACT statements. Parameters for these models appear on many statements including: MODELS, IMPACT, MOBILITY, and MATERIAL. The physical models can be grouped into five classes: mobility, recombination, carrier statistics, impact ionization, and tunneling. Chapter 3: “Physics”, Section 3.6: “Physical Models” contains details for each model.

All models with the exception of impact ionization are specified on the MODELS statement. Impact ionization is specified on the IMPACT statement. For example, the statement:

```
MODELS CONMOB FLDMOB SRH FERMIDIRAC
```

```
IMPACT SELB
```

specifies that the standard concentration dependent mobility, parallel field mobility, Shockley-Read-Hall recombination with fixed carrier lifetimes, Fermi Dirac statistics and Selberherr impact ionization models should be used.

ATLAS also provides an easy method for selecting the correct models for various technologies. The MOS, BIP, PROGRAM, and ERASE parameters for the MODELS statement configure a basic set of mobility, recombination, carrier statistics, and tunneling models. The MOS and BIP parameters enable the models for MOSFET and bipolar devices, while PROGRAM and ERASE enable the models for programming and erasing programmable devices. For example, the statement:

```
MODELS MOS PRINT
```

enables the CVT, SRH, and FERMIDIRAC models, while the statement:

```
MODELS BIPOLAR PRINT
```

enables the CONMOB, FLDMOB, CONSRH, AUGER, and BGN.

Note: The PRINT parameter lists to the run time output the models and parameters, which will be used during the simulation. This allows you to verify models and material parameters. We highly recommend that you include the PRINT parameter in the MODEL statement.

SSPD_Chapter 6_Part 7_Device Simulation2_concluded

SSPD_Chapter 6_Part 7_Device Simulation 2 is concluded in this module.

SSPD_Chapter 6_Part 7_Device Simulation 2_concluded

7.5. Bipolar Gummel Plot with Ic/Vce with constant Ib

```
go atlas

TITLE Bipolar Gummel plot and IC/VCE with constant IB

# Silvaco International 1992, 1993, 1994

mesh

x.m l=0 spacing=0.15

x.m l=0.8 spacing=0.15

x.m l=1.0 spacing=0.03

x.m l=1.5 spacing=0.12

x.m l=2.0 spacing=0.15

y.m l=0.0 spacing=0.006
```

```
y.m l=0.04 spacing=0.006

y.m l=0.06 spacing=0.005

y.m l=0.15 spacing=0.02

y.m l=0.30 spacing=0.02

y.m l=1.0 spacing=0.12

region num=1 silicon

electrode num=1 name=emitter left length=0.8

electrode num=2 name=base right length=0.5 y.max=0

electrode num=3 name=collector bottom

doping reg=1 uniform n.type conc=5e15 # background doping of the N-epi layer

doping reg=1 gauss n.type conc=1e18 peak=1.0 char=0.2 # N-type Collector implant

doping reg=1 gauss p.type conc=1e18 peak=0.05 junct=0.15 #P-type Base implant;

doping reg=1 gauss n.type conc=5e19 peak=0.0 junct=0.05 x.right=0.8 #N-type Emitter implant;

doping reg=1 gauss p.type conc=5e19 peak=0.0
```

```
char=0.08 x.left=1.5 # base contact P-type implant;

# set bipolar models

models conmob fldmob consrh auger print# consrh means concentration dependent lifetimes applicable in Silicon. Print has been explained in Section 7.4.

contact name=emitter n.poly surf.rec # surf.rec means surface recombination.

solve init

save outf=bjtex04_0.str

tonyplot bjtex04_0.str -set bjtex04_0.set

# Gummel plot

method newton autonr trap

solve vcollector=0.025

solve vcollector=0.1

solve vcollector=0.25 vstep=0.25 vfinal=2 name=collector

solve vbase=0.025

solve vbase=0.1
solve vbase=0.2
log outf=bjtex04_0.log
solve vbase=0.3 vstep=0.05 vfinal=1 name=base
tonyplot bjtex04_0.log -set bjtex04_0_log.set
#IC/VCE with constant IB
#ramp Vb
log off
solve init
solve vbase=0.025
solve vbase=0.05
solve vbase=0.1 vstep=0.1 vfinal=0.7 name=base
# switch to current boundary conditions
contact name=base current
# ramp IB and save solutions
solve ibase=1.e-6
save outf=bjtex04_1.str master
```

```
solve ibase=2.e-6

save outf=bjtex04_2.str master

solve ibase=3.e-6

save outf=bjtex04_3.str master

solve ibase=4.e-6

save outf=bjtex04_4.str master

solve ibase=5.e-6

save outf=bjtex04_5.str master

# load in each initial guess file and ramp VCE

load inf=bjtex04_1.str master

log outf=bjtex04_1.log

solve vcollector=0.0 vstep=0.25 vfinal=5.0
name=collector

load inf=bjtex04_2.str master

log outf=bjtex04_2.log

solve vcollector=0.0 vstep=0.25 vfinal=5.0
name=collector
```

```
load inf=bjtex04_3.str master

log outf=bjtex04_3.log

solve vcollector=0.0 vstep=0.25 vfinal=5.0
name=collector

load inf=bjtex04_4.str master

log outf=bjtex04_4.log

solve vcollector=0.0 vstep=0.25 vfinal=5.0
name=collector

load inf=bjtex04_5.str master

log outf=bjtex04_5.log

solve vcollector=0.0 vstep=0.25 vfinal=5.0
name=collector

# plot results

tonyplot -overlay bjtex04_1.log bjtex04_2.log
bjtex04_3.log bjtex04_4.log bjtex04_5.log -set
bjtex04_1_log.set

quit
```

7.5.1. Discussion on the Atlas Program used for BJT Device simulation and characterization.

Below in Figure 7.6 and Figure 7.7, the implant profile of the dopents are given. Figure 7.6 gives the absolute doping profile and Figure 7.7 gives the net doping profile after compensation.

From Figure 7.7 it is clear that EB Junction is at 0.05μm and BC Junction is at 0.15μm. Emitter is heavily doped at 5×10^{19} Phosphorous atoms per cc , Base is at 1×10^{18} Boron atoms per cc and Collector is at the background doping of 5×10^{15} Phosphorous atoms per cc. The base width is 0.10μm.

Figure 7.8 gives the mesh or the gridlines which cover the cross-sectional area of the device. As seen from the Mesh Diagram it is 0 to 1 μm in vertical or Y-direction and 0 to 2 μm in lateral direction or in X-direction.

As seen from Figure 7.8, finest mesh covers the encircled area and second finest mesh covers the bracketed are.

In Figure 7.9, 7.10 and 7.11 the Tony Plot of the structure of vertical NPN, I_C and I_B with respect to V_{BE} and family of output curves (I_C vs V_{CE}) are given respectively. Not in Figure 7.9 we see that ratio of I_C and I_B for constant V_{CE} is the short circuit d.c. current gain, β_F, which is poor at low currents due to recombination in the depletion layer and it is also poor at large currents due base conductivity modulation. It is optimum around 100 at $I_C = 0.5$ to 5mA. Emitter Injection Efficiency is degraded due to recombination in depletion layer and due to base conductivity modulation.

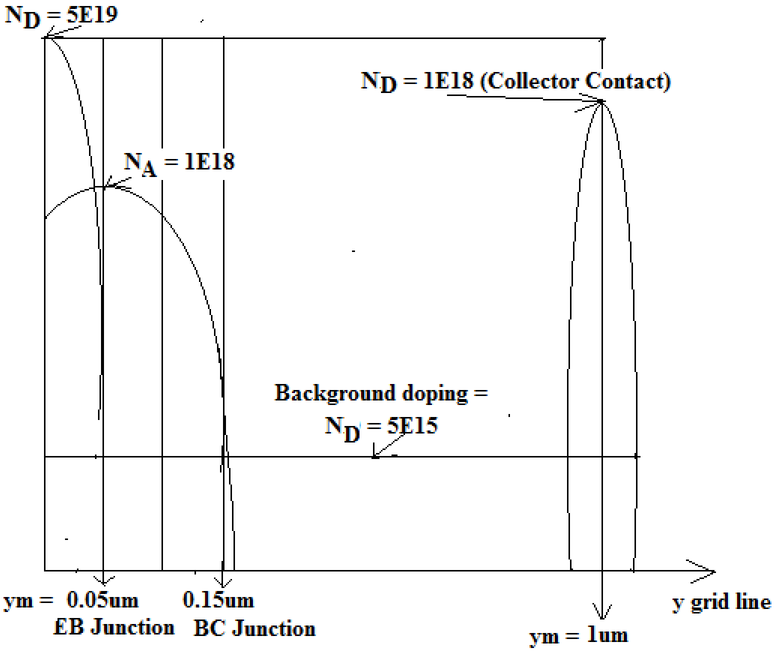


Figure 7.6. Absolute doping profile of implanted dopents.

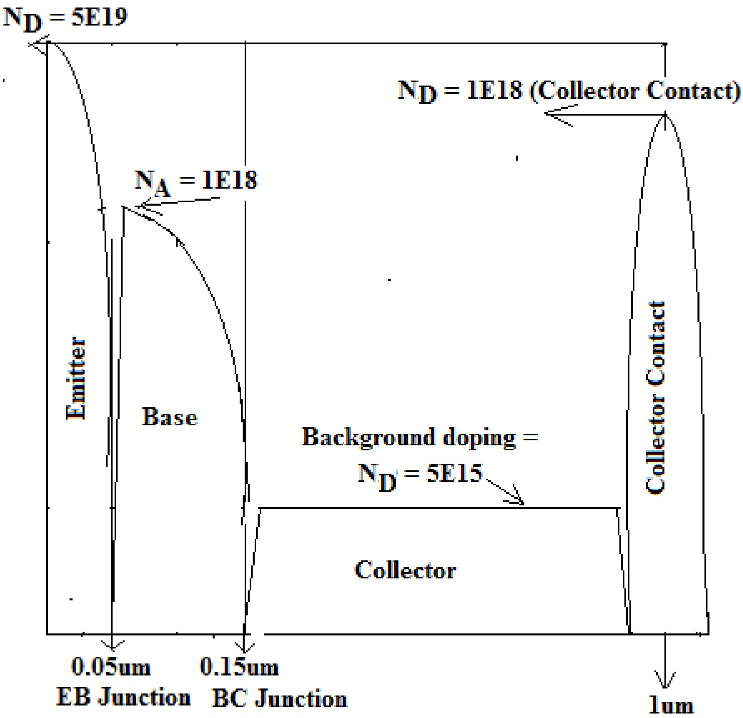


Figure 7.7 Net doping profile after compensation.

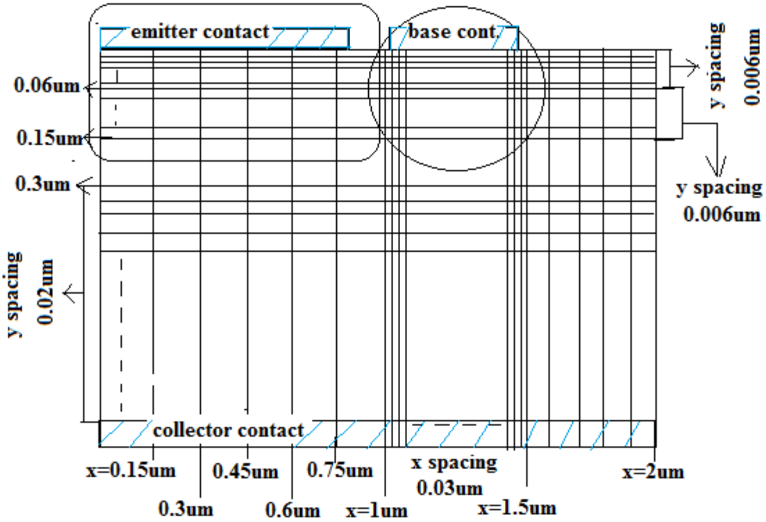


Figure 7.8. The mesh structure which covers the 2D cross-section of vertical NPN IC BJT.

Encircled area is the active part and is covered with the finest mesh. Bracketed area is covered by the next finest mesh.

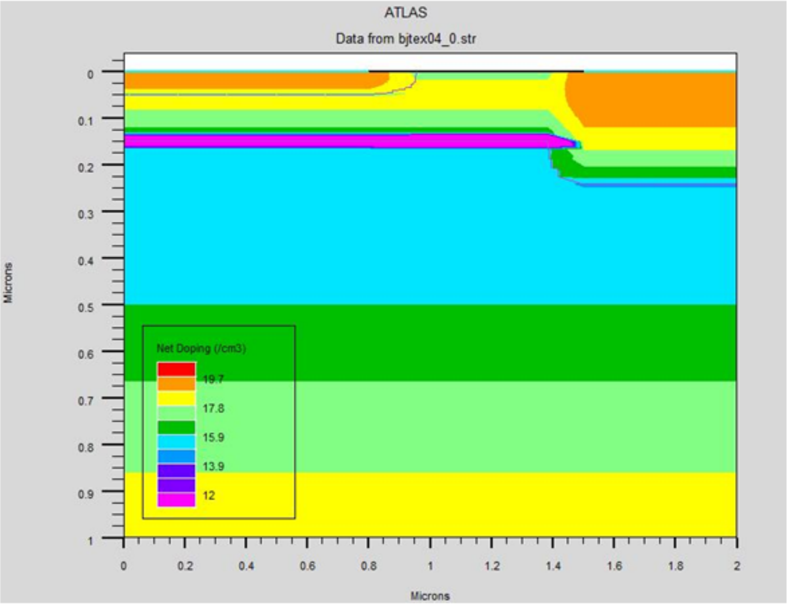


Figure 7.9. The structure generated by Tony Plot.

Figure 7.10. Variation of Base and Collector Current keeping Vce constant. The ratio of the two will give short circuit current gain . Note that at low currents β is unity and this gain plateaus at 5mA and at high current it again falls.

Figure 7.11. Output family of curves I C -V CE for the given transistor under constant current drive.

SSPD_Chapter 6_Part 7_Device Simulation2_Tony Plots

SSPD_Chapter 6_Part 7_Device Simulation 2_Tony Plots gives collector and base current generated and the output family of curves generated by BJT simulation.

SSPD_Chapter 6_Part 7_Device Simulation 2_concluded

7.5. Bipolar Gummel Plot with Ic/Vce with constant Ib

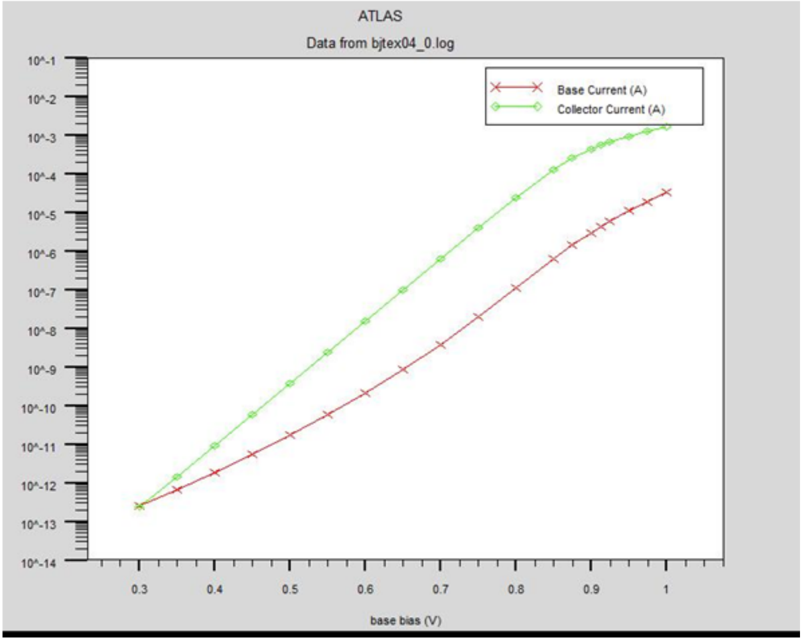


Figure 7.10. Variation of Base and Collector Current keeping Vce constant. The ratio of the

two will give short circuit current gain . Note that at low currents β is unity and this gain plateaus at 5mA and at high current it again falls.

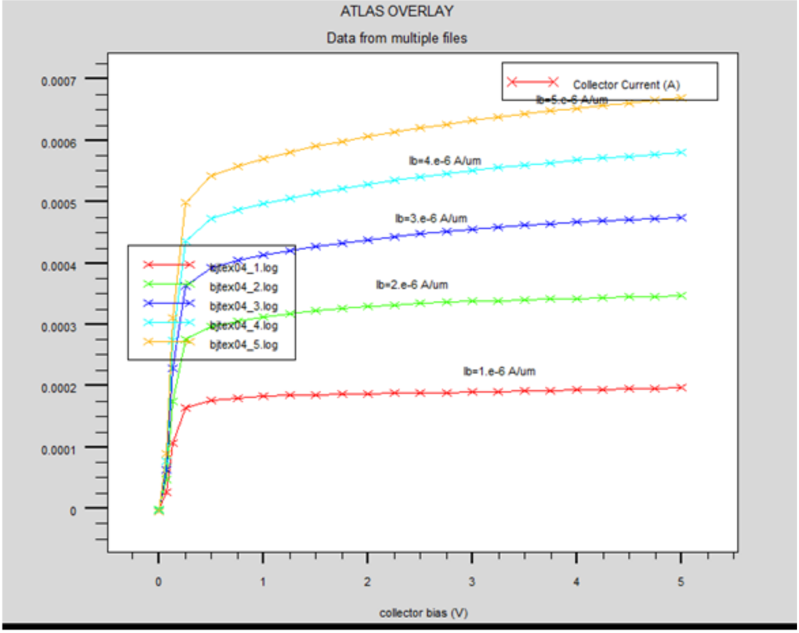


Figure 7.11. Output family of curves I C -V CE for the given transistor under constant current drive.

SSPD_Chapter 6_Part 7_Device Simlation 3
SSPD_Chapter 6_Part 7_Device Simulation 3 gives the simulation of Breakover phenomena in a Common Emitter BJT and its relationship to the fabrication parameter

SSPD_Chapter 7_Device Simulation 3.

7.6. Simulation of BV CEO in BJT and determining the process parameters which control this device d.c. parameter namely Break-over Voltage or Sustaining Voltage.

```
go atlas

TITLE Bipolar BVCEO simulation

# Silvaco International 1994

mesh

x.m l=0 spac=0.1

x.m l=2 spac=0.1

y.m l=0 spac=0.002

y.m l=1 spac=0.10

region num=1 silicon
```

```
electrode num=1 name=emitter left length=0.8

electrode num=2 name=base right length=0.5 y.max=0

electrode num=3 name=collector bottom

doping reg=1 uniform n.type conc=5e15

doping reg=1 gauss n.type conc=1e18 peak=1.0 char=0.2

doping reg=1 gauss p.type conc=1e18 peak=0.05 junct=0.15

doping reg=1 gauss n.type conc=5e19 peak=0.0 junct=0.05 x.right=0.8

doping reg=1 gauss p.type conc=5e19 peak=0.0 char=0.08 x.left=1.5

save outf=bjtex05_0.str

tonyplot bjtex05_0.str -set bjtex05_0.set

# set poly emitter

contact name=emitter n.poly surf.rec

material taun0=5e-6 taup0=5e-6

# set models
```

```
models bipolar print
impact selb
solve init
method newton trap
solve prev
solve vbase=0.025
solve vbase=0.05
solve vbase=0.2
contact name=base current
method newton trap ir.tol=1.e-20 ix.tol=1.e-20
solve ibase=3.e-15
log outf=bjtex05.log master
# ramp collector voltage
solve vcollector=0.25
solve vcollector=0.5
solve vcollector=1
solve vcollector=3

solve vcollector=5

solve vstep=0.5 vfinal=10 name=collector compl=5.e-11 e.comp=3

# plot results

tonyplot bjtex05.log -set bjtex05_log.set

quit
```

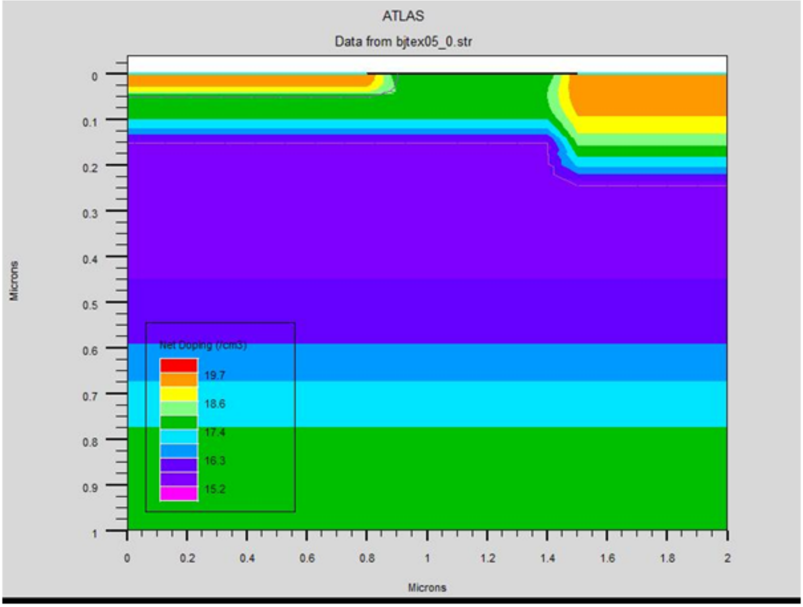


Figure 7.12. The cross sectional view of the BJT being simulated.

The Tony Plot of the output curve for zero base current is shown in Figure 7.13. A real BJT under similar condition shows a ‘S’ type Negative Impedance Region(NIR) which is not being shown in the Tony Plot. BVCEO is shown to be around 8V.

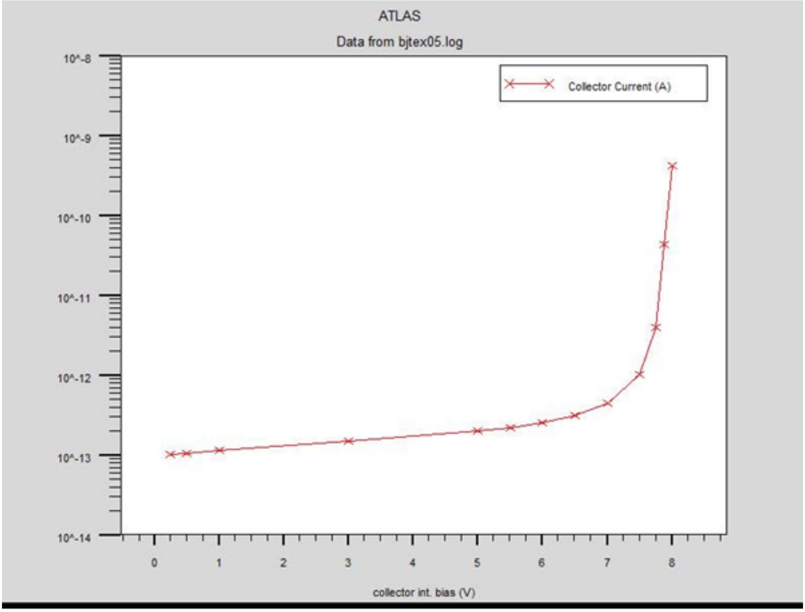


Figure 7.13. The output curve for base current zero.

7.6.1 Dependence of BV CEO on Base Width and Collector Base Doping.

As shown in Analog Electronics Lecture Notes,

$$BV_{CEO} = \frac{BV_{CBO}}{\sqrt[n]{1+\beta}}$$

7.3

As shown in SSPD notes for 2-sided step junction:

$$BV_{CBO} = \frac{\epsilon_s E_{crit}^2}{2q}(\frac{1}{N_D} + \frac{1}{N_A})$$

7.4

Where $E_{critical} = 3 \times 10^5$V/cm for Silicon.

A shown in Analog Electronics Lecture Notes,

$$\frac{I_C}{I_B} = \beta_F = \beta_{fo} = \frac{Q_{base}/\tau_{transit}}{Q_{base}/\tau_B} = \frac{\tau_B}{\tau_{transit}}$$

7.5

In Analog Electronics we have seen that transit time through the base:

$$\tau_{transit} = \frac{W^2}{2D_B}$$

7.6

Substituting Eq.7.6 in Eq.7.5 we get:

$$\beta_{fo} = \frac{\tau_B \times 2D_B}{W^2}$$

7.7

From Eq 7.3 it is clear the BV_{CEO} (the sustaining voltage of CE BJT) can be improved by increasing the Avalanche Breakdown Voltage of CB Junction and by decreasing short circuit current gain β_{fo} .

From Eq. 7.4, it is evident that breakdown can be increased by making CB junction lighter doped. From Eq.7.7, it is evident that short circuit current gain can be reduced by increasing base width (W).

7.6.2. Increase in Sustaining Voltage through making CB junction lighter doped.

CB junction is not a double –sided step junction instead it is linarly graded junction where the following formula is applicable :

$$W = \left(\frac{12\epsilon_S(\emptyset_{BO}+V_R}{qa}\right)^{1/3}$$

7.8

By equating $E_{max} = E_{critical}$ we obtain the CB breakdown voltage as:

$$BV_{CEO} = \left(\frac{E_{crit}}{2}\right)^{3/2}\left(\frac{12\epsilon_s}{qa}\right)^{1/2}$$

7.9

Here breakdown is critically dependent on ‘a’ = impurity gradient ($1/cm^4$). Therefore lowering of dopent concentration per se will not increase the sustaining voltage. We have to see how the change of dopent concentration effects the impurity gradient. Therefore lowering the dopent concentration is not tried out.

7.6.3. Increase in sustaining voltage by increasing the Base Width (W).

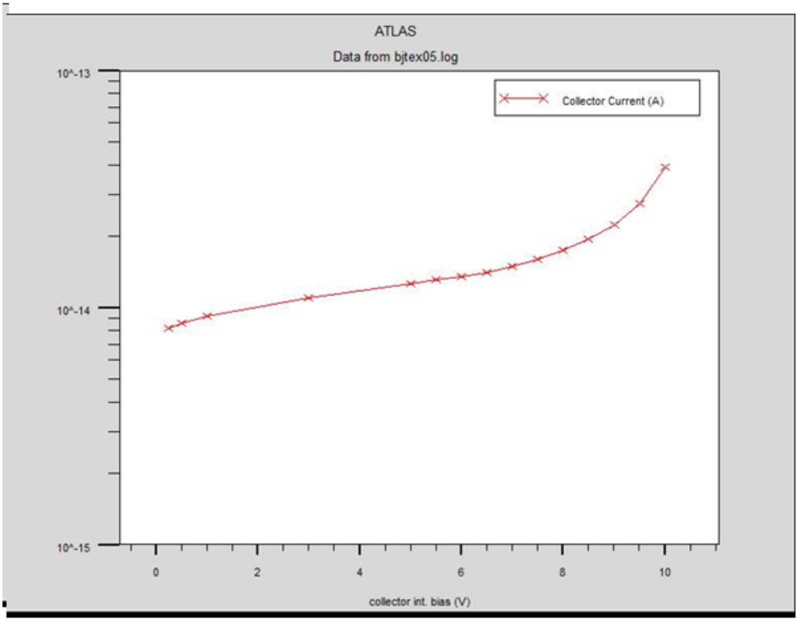


Figure 7.14. The Output Curve of BJT with Base Current zero and with Base Width increased.

In this case Base Width is increased from 1 micron to 1.2 micron by shifting BC Junction to 0.16micron and decreasing EB Junction to 0.04 micron. This has very clearly increased the sustaining voltage from

8V to 10V.

```
go atlas
TITLE Bipolar BVCEO simulation
# Silvaco International 1994
mesh
x.m l=0 spac=0.1
x.m l=2 spac=0.1
y.m l=0 spac=0.002
y.m l=1 spac=0.10
region num=1 silicon
electrode num=1 name=emitter left length=0.8
electrode num=2 name=base right length=0.5 y.max=0
electrode num=3 name=collector bottom
doping reg=1 uniform n.type conc=5e15
doping reg=1 gauss n.type conc=1e18 peak=1.0 char=0.2
doping reg=1 gauss p.type conc=1e18 peak=0.05
junct=0.16

doping reg=1 gauss n.type conc=5e19 peak=0.0
junct=0.04 x.right=0.8

doping reg=1 gauss p.type conc=5e19 peak=0.0
char=0.08 x.left=1.5

save outf=bjtex05_0.str

tonyplot bjtex05_0.str -set bjtex05_0.set

# set poly emitter

contact name=emitter n.poly surf.rec

material taun0=5e-6 taup0=5e-6

# set models

models bipolar print

impact selb

solve init

method newton trap

solve prev

solve vbase=0.025
solve vbase=0.05

solve vbase=0.2

contact name=base current

method newton trap ir.tol=1.e-20 ix.tol=1.e-20

solve ibase=3.e-15

log outf=bjtex05.log master

# ramp collector voltage

solve vcollector=0.25

solve vcollector=0.5

solve vcollector=1

solve vcollector=3

solve vcollector=5

solve vstep=0.5 vfinal=10 name=collector compl=5.e-11 e.comp=3

# plot results

tonyplot bjtex05.log -set bjtex05_log.set

quit
```

SSPD_Chapter 6_Part 7_Device Simlation 3_conclsions
SSPD_Chapter 6_Part 7_device simulation 3_ points out the inadequecies of ATLAS simulatin software.

SSPD_Chapter 6_Part 7_Device Simulation 3_Conclusions.

7.6.3 Family of Output Curves of a real life BJT under constant base current drive and under constant base voltage drive.

The family of curves which are given in the books are constant current drive Output Curves and are shown in the bottom half of Figure 7.16. In the upper half the constant voltage drive family of Output Curves are shown. Constant Current and Constant Voltage Drive CE BJT circuit configurations are shown in Figure 7.15.a and 7.15.b..

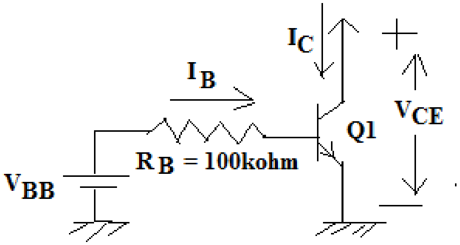


Figure 7.15a. Determination of Output Family of Curves under constant Base Current Drive.

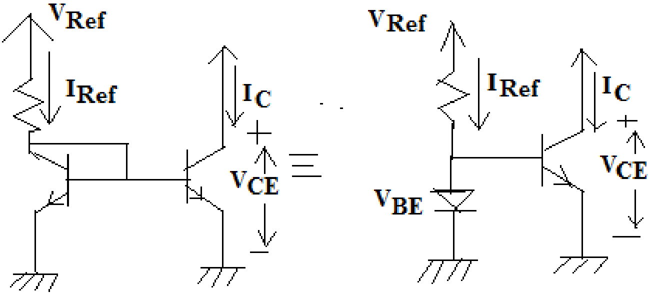
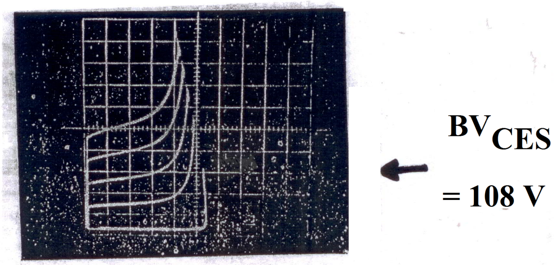
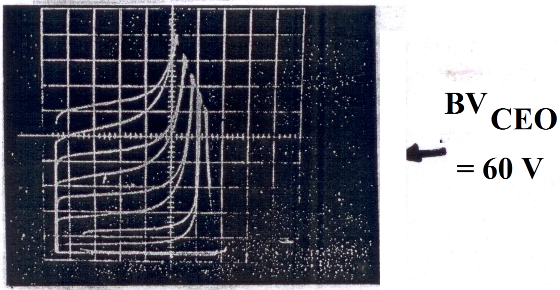


Figure 7.15b. Determination of Output Family of Curves under constant Base Voltage Drive. Constant Base Voltage Drive is achieved by Current Mirror Configuration of CE BJT.

A battery voltage with a large source resistance say of 100kohm is equivalent to Constant Current Source determined by Norton Equivalent Circuit. A battery voltage with a small source resistance say of 1ohm is equivalent to Constant Voltage Source determined by Thevenin Equivalent Circuit. The circuit in Figure 7.15.a is like CE BJT driven by constant base current. The circuit in Figure 7.15.b is like CE BJT driven by constant base voltage. Figure 7.15.b is Current Mirror Current Source.



X—20 V/cm, Y—50 μA/cm
Fig 9 I_C—V_{CE} output characteristic of a current mirror current source



X—10V/cm, Y—50 μA/cm
Fig 8 I_C—V_{CE} output characteristic of CE BJT 2N 2219 A

Figure 7.16.a. Family of Output Curves of CE BJT under constant Base Voltage Drive. Here we obtain BV_{CES} = 108V. This parameter means Break-over voltage at CE with Base shorted. When Base is driven by constant voltage then incrementally it means Base shorted.
Figure 7.16.b. Family of Output Curves of CE BJT under constant Base Current Drive. Here we obtain BV_{CEO} = 60V. This parameter means Break-over voltage at CE with Base open. When Base is driven by constant current then incrementally it means Base open.

The circuit in Figure 7.15.b is equivalent to CE BJT driven by constant base voltage. Figure 7.15.b is Current Mirror Current Source. The collector current in the second BJT is constraint to be equal to I_{Ref} in first BJT. Hence it is called Current Mirror Current

Source. Since I_{Ref} is the collector Current of a strong feed-back system(in this case it is a strong feed back CE BJT- voltage-shunt feed back) therefore I_{Ref} doesnot change with aging, with temperature changes or with replacement of the active device. Hence collector current in the second transistor is kept constant under all circumstances therefore second transistor behaves like a Current Source.

As we see in Figure 7.16.b. for $I_B = 0$ μA, the output curve breaksover at 60V whereas in Figure 7.16.a. for $I_B = 0$μA the output curve breaks over at 108 V. Also in constant current drive, as shown in Figure 7.16.b, at the breakover there is S-type Negative Impedance Region. We donot see this S-type NIR in the tony plot of the third device simulation.

The third device simulation does not exhibit the improvement in breakover when the input drive is changed from constant current to constant voltage drive. The compact model used in simulation software therefore needs to be improved.

SSPD_Chapter 6_Part 7_ Diode Tutorial
SSPD_Chapter 6_Part 7_Tutorial on Diode gives problem of design and simulation on Varctor Diode, Ideality Factor and Tunnel Diode.

SSPD_Chapter 6_Part 7_ Diode simulation Tutorial 1

1. Design a Varactor Diode and verify it by Device Simulatin program of ATLAS.[Hint:Step Junction Diode doping parameter m = 0, linearly graded junction diode m =1 and in Varactor Diode m = -3/2. Doping Profile analytic relation of Hyper-abrupt junction is given by the following expression:

$$N_D = B(\frac{x-x_0}{x_0})^m U(x - x_0) where\, m = \frac{-3}{2}$$

;

depletion width varies as the square root of the reverse bias voltage in step junction diode, varies as cube root of the reverse bias voltage in linearly graded junction diode and varies as the square of the reverse bias voltage in hyperabrupt junction diode where

m = -3/2. The varactor diode is used to assemble Voltage Controlled Oscillator (VCO) which is used in generating Frequency Modulated(FM) Carriers].

1. Design a step junction diode and study the forward bias I-V characteristics. From I-V characteristics extract the ideality factor for different ranges of current. See what device fabrication parameter control the ideality factor. [Hint: In Shockley Diode Equation we have the two basic assumptions: a. Bulk is quasi-neutral region. b. there is no recombination in the depletion region. This leads to :

$$V_D\,(diode\ terminal\ voltage) = \ V_J = Junction\ Voltage.$$

But at low currents (below 0.5mA) assumption (b) is violated and at high currents i.e at currents greater than 5mA assumption (a) is violated. In both cases :

$$\begin{aligned} V_D\,(diode\ terminal\ voltage) &= \ V_J(Junction\ Voltage) + V_{Bn} + V_{Bp} \\ &= 2V_J \end{aligned}$$

Hence in diode equation :

$$I_D = \ I_{D0}\left[Exp\left(\frac{V_J}{V_T}\right) - \ 1\right]$$

Under two conditions, when I_D is less tthan 0.5mA and greater than 5mA, $V_J = V_D/2$ and therefore real diode equation under these two conditions is:

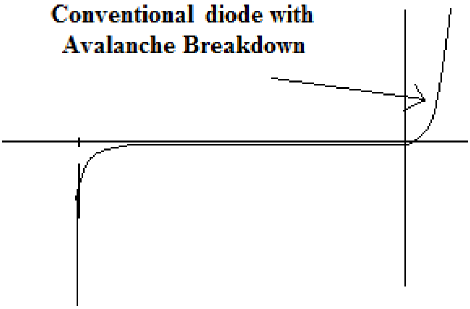
$$I_D = I_{D0}\left[Exp\left(\frac{V_D}{2V_T}\right) - 1\right]$$

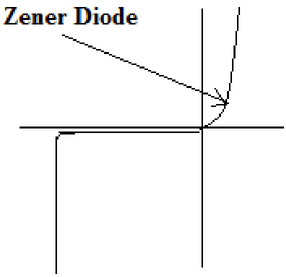
Therefore under these two conditions we have an Ideality Factor of η = 2 and real diode equation is given by:

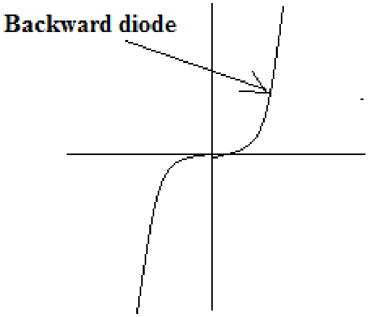
$$I_D = I_{D0}\left[Exp\left(\frac{V_D}{\eta V_T}\right) - 1\right]$$

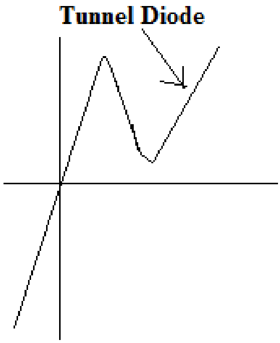
]

1. Design a tunnel diode and simulate its I-V characteristics. As the doping concentration is raised the I-V characteristics are changed from conventional diode to Zener Diode to Backward Diode to Tunnel Diode. These changes in I-V characteristics can be tested by scaling up the doping.









Type	Anode doping	Cathode doping
Conventionaldiode	10^{16} dopent atoms/cc	10^{16} dopent atoms/cc
Zener Diode	10^{19} dopent atoms/cc	10^{19} dopent atoms/cc
Backward Diode	10^{20} dopent atoms/cc	10^{20} dopent atoms/cc
Tunnel Diode	10^{21} dopent atoms/cc	10^{21} dopent atoms/cc

SSSPD_Chapter 6_Part 7_Introduction to ATHENA_Process Simulator
SSPD_Chapter 6_Part 7_Introduction to ATHENA. This section describes how to use DECKBUILD's Commands Menu to create a typical ATHENA input file. The goal is not to design a real process sequence but to demonstrate the use of specific ATHENA statements and parameters as well as some DeckBuild features to create a realistic input file. You can find many realistic process files among the examples and use them as a starting point on your process simulator.

SSPD_Chapter 6_Part 7_Introduction to ATHENA1(Process Simulator)

7.7. Background of ATHENA (Process Simulator)

The ATHENA Two-Dimensional Process Simulation Framework is a comprehensive software tool for modeling semiconductor fabrication processes. ATHENA provides facilities to perform efficient simulation analysis that substitutes for costly real world experimentation. ATHENA combines high temperature process modeling such as impurity diffusion and oxidation, topography simulation, and lithography simulation in a single, easy to use framework.

ATHENA is a simulator that provides general capabilities for numerical, physically-based, two-dimensional simulation of semiconductor processing. ATHENA has a modular architecture with the following licensable tools and extensions:

- ATHENA: This tool performs structure initialization and manipulation, and provides basic deposition and etch facilities

- SSUPREM4: This tool is used in the design, analysis, and optimization of silicon semiconductor structures. It simulates silicon processing steps such as ion implantation, diffusion and oxidation.

- ELITE: This tool is a general purpose 2D topography simulator that accurately describes a wide range of deposition, etch and reflow processes used in modern IC technologies.

- OPTOLITH: This tool performs general optical lithography simulation including 2D aerial imaging, non-planar photoresist exposure, and post exposure bake and development.

Table 7.4: Athena Features and Capabilities

Features	Capabilities
Bake	• Time and temperature

	bake specification.• Models photoresist material flow.• Models photo-active compound diffusion.
C-Intepreter	• Allows user-defined models for implant damage, Monte Carlo plasma etching and diffusion in SiGeC.
CMP	• Models Chemical Mechanical Polishing.• Hard and soft models or a combination of both.• Includes isotropical etch component.
Deposition	• Conformal deposition model.• Hemispherical, planetary, and conical metallization models.• Unidirectional or dual directional deposition models.• CVD model.• Surface diffusion/ migration effects.• Ballistic deposition models including atomistic positioning effects.• User-definable models.• Default deposition machine

	definitions.
Mask Development	• Five different photoresist development models.
Diffusion	• Impurity diffusion in general 2D structures including diffusion in all material layers.• Fully coupled point defect diffusion model.• Oxidation enhanced/ retarded diffusion effects.• Rapid thermal annealing.• Models simultaneous material reflow and impurity diffusion.• Impurity diffusion in polysilicon accounting for grain and grain boundary components.
Epitaxy	• 2D epitaxy simulation including auto doping.
Etch	• Extensive geometric etch capability.• Wet etching with isotropic profile advance.• RIE model that combines isotropic and directional etch components.• Microloading effects.• Angle dependence of

	etchant source.• Default etch machine definitions.• Monte Carlo plasma etching.• Dopant enhanced etching.
Exposure	• Model is based on the Beam Propagation Method simulating reflections and diffraction effects in non-planar structures with capability to take into account local modification of material optical properties the absorbed dose.• Defocus and large numerical aperture effects.
Imaging	• Two dimensional, large numerical aperture, aerial image formation.• Up to 9th order imaging system aberrations.• Extensive source and pupil plane filtering for enhanced aerial images.• Full phase shift and transmittance variation mask capabilities.
Implantation	• Experimentally verified Pearson and dual Pearson analytical models.•

	Extended low energy and high energy implant parameter tables.• Binary Collision Approximation Monte Carlo calculations for crystalline and amorphous materials.• Universal tilt and rotation capability for both analytic and Monte Carlo calculations.
Oxidation	• Compressible and viscous stress dependent models.• Separate rate coefficients for silicon and polysilicon materials.• HCL and pressure-enhanced oxidation models.• Impurity concentration dependent effects.• Ability to simulate the oxidation of structures with deep trenches, undercuts, and ONO layers.• Accurate models for the simultaneous oxidation and lifting of polysilicon regions.
Silicidation	• Models for titanium, tungsten, cobalt, and

platinum silicides.• Experimentally verified growth rates.• Reactions and boundary motion on silicide/metal and silicide/silicon interfaces.• Accurate material consumption model.

In Athena we are doing physically based simulation.The Value Of Physically-Based Simulation:

Physically-based process simulators predict the structures that result from specified process sequences. This is done by solving systems of equations that describe the physics and chemistry of semiconductor processes.

Physically-based simulation provides three major advantages: it is predictive, it provides insight, and it captures theoretical knowledge in a way that makes this knowledge available to non-experts.

Physically-based simulation is different from empirical modeling. The goal of empirical modeling is to obtain analytic formulae that approximate existing data with accuracy and minimum complexity. Empirical models provide efficient approximation and interpolation. Empirical models, however, doesn’t provide insight, predictive capabilities, or capture theoretical knowledge. Physically-based simulation is an alternative to experiments as a source of data. Empirical modeling can provide compact representations of data from either source.

Physically-based simulation has become very important for two reasons. One, it’s almost always much quicker and cheaper than performing experiments. Two, it provides information that is difficult or impossible to measure.

Physically-based simulation has two drawbacks: you must incorporate all the relevant physics and chemistry into a simulator and numerical procedures, and you must be implement them to solve the associated equations. But these tasks have been taken care of for ATHENA users.

Physically-based process simulation tools users must specify the problem to be simulated. ATHENA users specify the problem by defining the following:

• The initial geometry of the structure to be simulated. • The sequence of process steps (e.g., implantation, etching, diffusion, exposure) that are to be simulated. • The physical models to be used.

7.7.1: Using ATHENA with other SILVACO Software

ATHENA is normally used in conjunction with the

VWF INTERACTIVE TOOLS. These tools include DECKBUILD, TONYPLOT, DEVEDIT, MASKVIEWS and OPTIMIZER. DECKBUILD provides an interactive run time environment. TONYPLOT supplies scientific visualization capabilities. DEVEDIT is an interactive tool for structure and mesh specification and refinement, and MASKVIEWS is an IC Layout Editor. The OPTIMIZER supports black box optimization across multiple simulators.

ATHENA is also frequently used in conjunction with the ATLAS device simulator. ATHENA predicts the physical structures that result from processing. These physical structures are used as input by ATLAS, which then predicts the electrical characteristics associated with specified bias conditions. *Using ATHENA and ATLAS makes it easy to determine the impact of process parameters on device characteristics.*

7.7.1. Getting started with ATHENA.

To start ATHENA under DECKBUILD in interactive mode, enter the following UNIX command:

```
deckbuild -an
```

After a short delay, the Main Deckbuild Window will appear. The lower text window of this window will contain the ATHENA logo and version number, a list of available modules, and a command prompt.

ATHENA is now ready to run. To become familiar with the mechanics of running ATHENA under DECKBUILD, load and run some of the ATHENA standard examples.

7.7.2.: Creatinga Device Structure Using ATHENA

7.7.2.1: Procedure Overview

ATHENA is designed as a process simulation framework. The framework includes simulator independent operations and simulator specific functions that simulate different process steps (e.g., implant, RIE, or photoresist exposure). This section describes ATHENA input/output and the following basic operations for creating an input file:

- Developing a good simulation grid
- Performing conformal deposition
- Performing geometric etches
- Structure manipulation
- Saving and loading structure information
- Interfacing with device simulators
- Using different VWF INTERACTIVE TOOLS

7.7.2.2: Creating An Initial Structure

This section will describe how to use DECKBUILD’s Commands menu to create a typical ATHENA input file. The goal of this section is not to design a real process sequence, but to demonstrate the use of specific ATHENA statements and parameters, as well as some DECKBUILD features, to create a realistic input file. You can find many realistic process input files among the examples and use them as a starting point in your process simulation.

Once DECKBUILD is running and the current simulator is set to ATHENA (see the VWF AUTOMATION, CALIBRATION, AND PRODUCTION TOOLS USER ’S MANUAL for more information), open and pin the Commands menu as shown in Table 7.5.. Then, select Mesh Define.... and the ATHENA Mesh Define Menu will appear. We recommend that you pin this popup because it will be used often in designing an initial mesh.

Table 7.5 Commands Menu

Commands
Mesh define
Mesh initialize

Adaptive meshing
Process ►
Structure ►
File I/O
Notes……
Models…..
Templates…….
Extract
Parse Deck

Defining Initial Rectangular Grid.

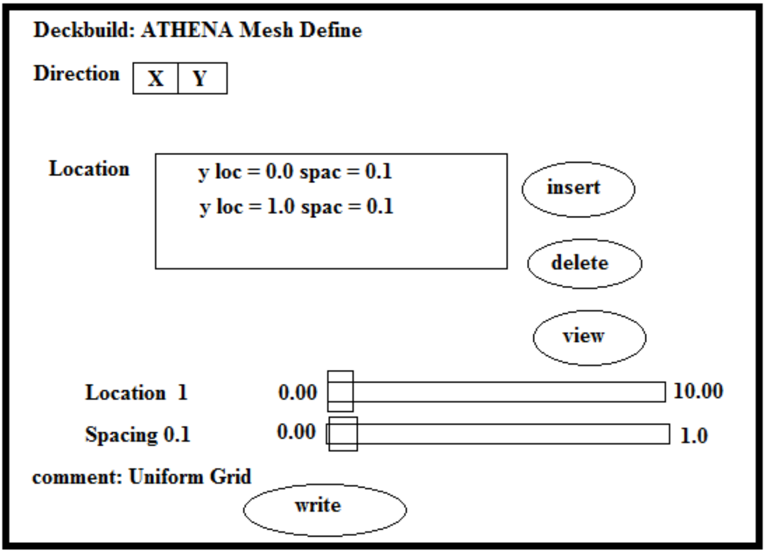
Now, you can specify the initial rectangular grid. The correct specification of a grid is critical in process simulation. The number of nodes in the grid Np has a direct influence on simulation accuracy and time. A finer grid should exist in those areas of the simulation structure where ion implantation will occur, where p-n junction will be formed, or where optical illumination will change photoactive component concentration. The number of arithmetic operations necessary to achieve a solution for processes simulated, using the finite element analysis method could be estimated as $(Np)_a$, where a is of order 1.5 - 2.0. Therefore, to maintain the simulation time within reasonable bounds, the fine grid should not be allowed to spill over into unnecessary regions. The maximum number of grid nodes is 20,000 for ATHENA simulations, but most practical simulations use far fewer nodes than this limit.

To create a simple uniform grid in a rectangular 1 µm by 1 µm simulation area, click on the Location field and enter a value of 0.0. Then, click on the Spacing field and enter a value of 0.10. Then, click on the Insert button and the line parameters will appear in the scrolling list.

Note: ATHENA coordinate system has positive x axis pointed to the right along the structure surface and positive y axis pointed down to the depth of the structure.

In the same way, set the location of a second X line to 1.0 with a spacing of 0.1. You can either set the values by dragging a slider or by entering a number directly.

Now, select the Y direction and set the lines with the same values as the X direction. You can now add the comments at the Comment line. The ATHENA Mesh Define menu should appear as shown in Figure 7.17.



Deckbuild: ATHENA Mesh Define

Direction X Y

Location

y loc = 0.0 spac = 0.1

y loc = 1.0 spac = 0.1

Location 1 0.00 10.00

Spacing 0.1 0.00 1.0

comment: Uniform Grid

Figure 7.17. ATHENA Mesh Define Menu

You can now write the menu-prepared mesh information into the input file. But first, preview the rectangular grid by selecting the View... button and the View Grid window (Figure 7.18) will appear. Notice that vertical and horizontal grid lines are distributed uniformly, and the 121 points and the 200 triangles will be generated.

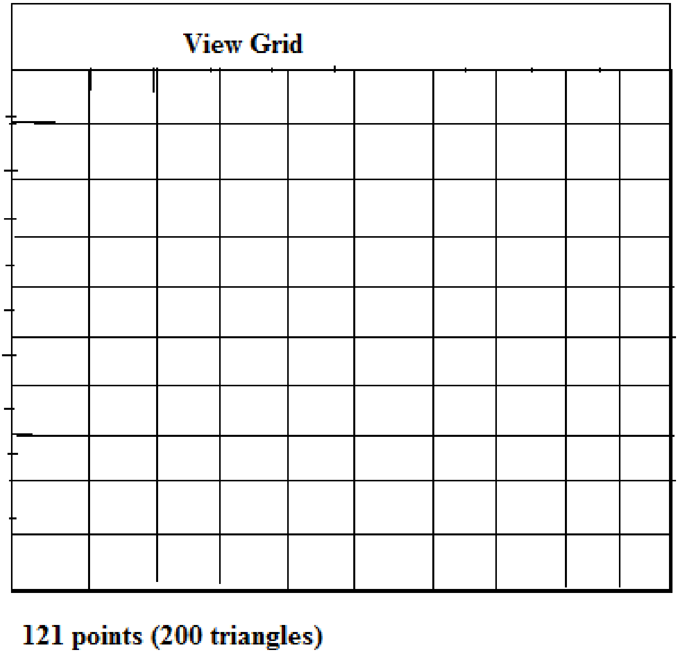


Figure 7.18. View Grid Window.

A uniform grid such as the one shown in Figure 7.18 is inefficient for performing complex simulations. Therefore, the grid must be improved. First, make a better grid in the y-direction. Usually, it's necessary to get better resolution for the depth profile after the ion implantation step. When adaptive gridding capability isn't used, apply preliminary knowledge of the process you are going to simulate.

Suppose you want to perform a 60 keV boron implant so that the implant peak would be around 0.2 μm. It is reasonable to make a finer grid at this depth. To achieve this, simply add one more Y-line by setting the Location to 0.2 and the Spacing to

0.02. The new rectangular grid (Figure 7.19) will now appear. Notice the number of points and triangles have increased to 231 and 400 respectively.

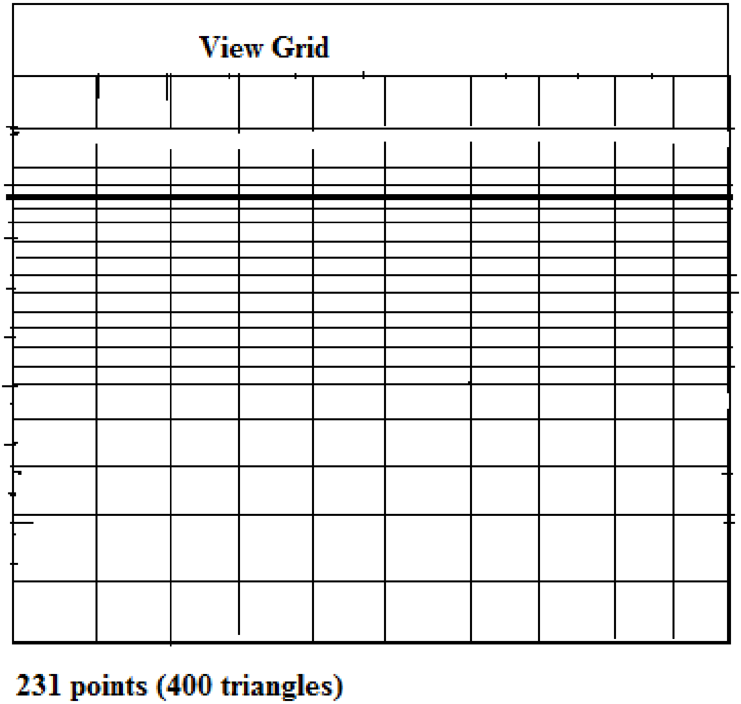


Figure 7.19. New Rectangular Grid

Finally, write the Mesh Define information to the file by pressing the Write button. A set of lines like these will appear:

GO ATHENA

# NON-UNIFORM GRID

```
LINE X LOC=0.00 SPAC=0.1

LINE X LOC=0.3 SPAC=0.02

LINE X LOC=1 SPAC=0.1

LINE Y LOC=0.00 SPAC=0.03

LINE Y LOC=0.2 SPAC=0.02

LINE Y LOC=1 SPAC=0.1
```

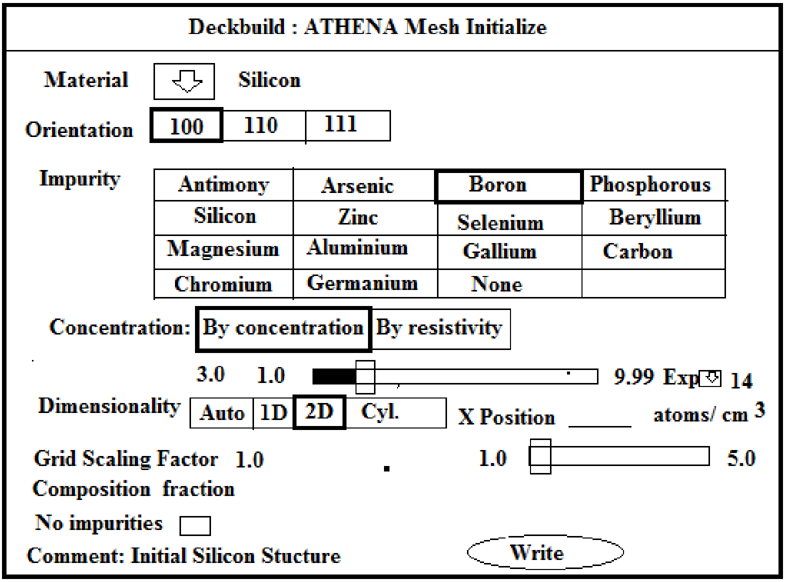
The first line (GO ATHENA) is called an autointerface statement and tells DECKBUILD that the following file should be run by ATHENA

The grid in ATHENA consists of points connected to form a number of triangles. Each point has one or more nodes associated with it. A point within a material region has one node, while a point which belongs to several regions has several nodes. A node represents the solution (e.g., doping concentration) in a particular material region at the point. For example, a given node may represent solution values in silicon at a point with coordinates (0.0, 0.0); an entirely different node may represent solution values in oxide at the same point (0.0, 0.0).

So, the previous INIT statement creates the <100> silicon region of 1.0 µm x 1.0 µm size, which is uniformly doped with boron concentration of 3e14 atom/cm3. This simulation structure is ready for any process step (e.g., implant, diffusion, Reactive Ion Etching). Before discussing the simulation of physical processing using SSUPREM4, ELITE or OPTOLITH modules, it’s important to discuss structure manipulation statements that can precede or alternate with physical process steps.

7.7.2.3. Defining the initial substrate.

The LINE statements specified by the Mesh Define menu set only the rectangular base for the ATHENA simulation structure. The next step is the initialization of the substrate region with its points, nodes, triangles, background doping, substrate orientation, and some additional parameters. To initialize the simulation structure, select ATHENA Command Menu®Mesh Initialize... and the Mesh Initialize Menu will appear (see Figure 7-20)



Deckbuild : ATHENA Mesh Initialize

Material Silicon

Orientation **100** 110 111

Impurity

Antimony	Arsenic	**Boron**	Phosphorous
Silicon	Zinc	Selenium	Beryllium
Magnesium	Aluminium	Gallium	Carbon
Chromium	Germanium	None	

Concentration: **By concentration** By resistivity

3.0 1.0 9.99 Exp 14

Dimensionality Auto 1D **2D** Cyl. X Position _____ atoms/ cm ^{3}

Grid Scaling Factor 1.0 1.0 5.0

Composition fraction

No impurities ☐

Comment: Initial Silicon Stucture

Figure 7.20. Mesh Initialize Menu.

Background doping can be set by clicking on the desired impurity box (e.g., Boron). The background impurity concentration specification will then become active. If the None box is checked, the concentration information will become inactive and will appear grayed out from the rest of the menu. Select the desired concentration using the slider (e.g., 3.0) and select an exponent from the **Exp**: menu (e.g., 14). This will give a background concentration of 3.0e14 atom/cm3. You can set background concentration using the By **Resistivity** specification in Ohm•cm. For this tutorial, check the 2D box in the **Dimensionality** field. This will run the simulation in a two-dimensional calculation.

Note: Two-dimensional mode is used in this tutorial to demonstrate 2D grid generation and manipulation. In most cases, however, it is unnecessary to change the Auto default in the Dimensionality item of the Mesh Initialize menu. ATHENA will begin in 1D and will automatically switch to 2D mode at the first statement, which disrupts the lateral uniformity of the device structure. This generally results in massive savings of computation time.

You can now write the mesh initialization information into the file by pressing the Write button. The following two lines will appear in the Deckbuild Text Subwindow:

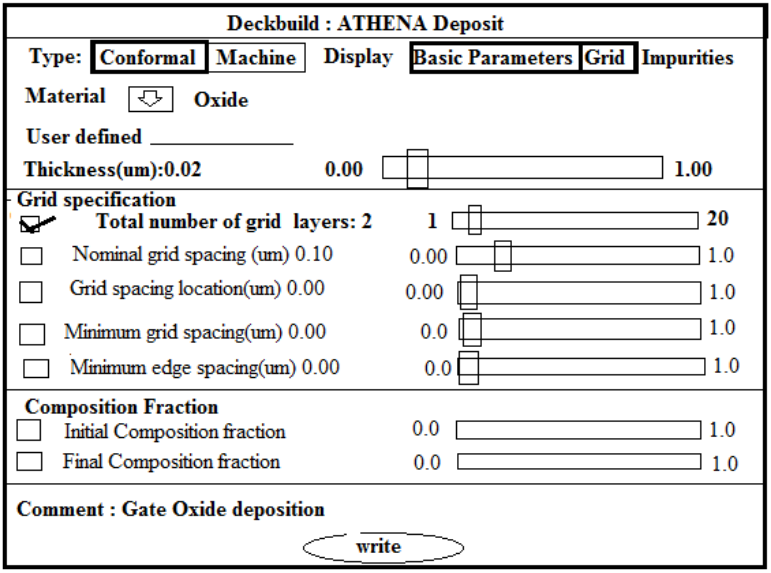
```
# INITIAL SILICON STRUCTURE

INIT SILICON C.BORON=3.0E14
ORIENTATION=100 TWO.D
```

So, the previous INIT statement creates the <100> silicon region of 1.0 mm x 1.0 mm size, which is uniformly doped with boron concentration of 3e14 atom/cm3. This simulation structure is ready for any process step (e.g., implant, diffusion, Reactive Ion Etching). Before discussing the simulation of physical processing using SSUPREM4, ELITE or OPTOLITH modules, it’s important to discuss structure manipulation statements that can precede or alternate with physical process steps.

7.7.2.4. Simple Film Depositions

Conformal deposition can be used to generate multi-layered structures. Conformal deposition is the simplest deposit model and can be used in all cases when the exact shape of the deposited layer is not critical. Conformal deposition can also be used in place of oxidation of planar or quasi-planar semiconductor regions when doping redistribution during the oxidation process is negligible.

To set the conformal deposition step, select the menu items Process→Deposit→Deposit... from the Commands menu in DECKBUILD and the ATHENA Deposit Menu (Figure 7-21) will appear.



Deckbuild : ATHENA Deposit

Type: Conformal | Machine   Display: Basic Parameters | Grid | Impurities

Material: Oxide

User defined ____________

Thickness(um):0.02   0.00 – 1.00

Grid specification

☑ Total number of grid layers: 2   1 – 20

☐ Nominal grid spacing (um) 0.10   0.00 – 1.0

☐ Grid spacing location(um) 0.00   0.00 – 1.0

☐ Minimum grid spacing(um) 0.00   0.0 – 1.0

☐ Minimum edge spacing(um) 0.00   0.0 – 1.0

Composition Fraction

☐ Initial Composition fraction   0.0 – 1.0

☐ Final Composition fraction   0.0 – 1.0

Comment : Gate Oxide deposition

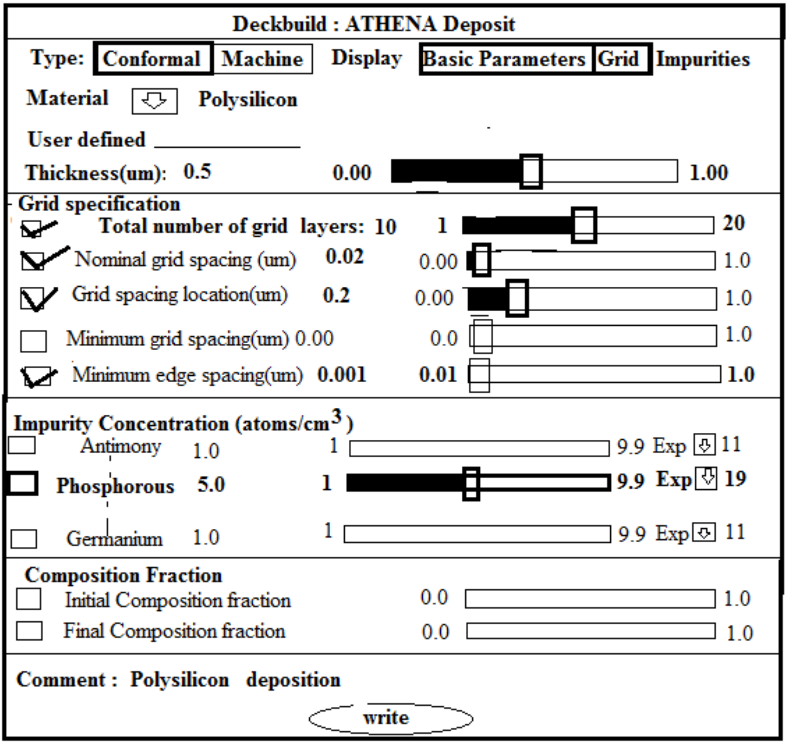
Figure 7.21. ATHENA Deposit Menu.

As shown, Conformal Deposition is the default. If it is known that the oxide layer thickness grown in a process is 200 Angstroms, you can substitute this with conformal oxide deposition. Select Oxide from the Material menu and set its thickness to 0.02 µm. It is always useful to set several grid layers in a deposited layer. In this case, at least two grid layers are needed to simulate impurity transport through the oxide layer. In some other cases (e.g., photoresist deposition over a non-planar structure), a sufficiently fine grid is needed to accurately simulate processes within the deposited layer. There are also situations (e.g., spacer formation) when several grid layers in a deposited material region are needed to properly represent the geometrical shape of the region.

The grid in the deposited layer is controlled by **Grid Specification** parameters in the ATHENA Deposit Menu. Set the **Total number of grid layers** to 2, add a **Comment**, and click on the **Write** button. The following lines will then appear in the Deckbuild Text Subwindow:

```
# GATE OXIDE DEPOSITION

DEPOSIT OXIDE THICK=0.02 DIVISIONS=2
```

The next step will be to deposit a phosphorus doped polysilicon layer of 0.5µm thickness. Select the material Polysilicon, and set the thickness to 0.5. To add doping, select the Impurities box. The Impurity Concentration section will be immediately added to the ATHENA Deposit Menu (See Figure 7-22).


Deckbuild : ATHENA Deposit
Type: Conformal Machine
Display Basic Parameters Grid Impurities
write

Material Polysilicon

User defined

Thickness(um): 0.5 0.00 1.00

Grid specification

☑ Total number of grid layers: 10 1 20

☑ Nominal grid spacing (um) 0.02 0.00 1.0

☑ Grid spacing location(um) 0.2 0.00 1.0

☐ Minimum grid spacing(um) 0.00 0.0 1.0

☑ Minimum edge spacing(um) 0.001 0.01 1.0

Impurity Concentration (atoms/cm^3)

☐ Antimony 1.0 1 9.9 Exp 11

☐ Phosphorous 5.0 1 9.9 Exp 19

☐ Germanium 1.0 1 9.9 Exp 11

Composition Fraction

☐ Initial Composition fraction 0.0 1.0

☐ Final Composition fraction 0.0 1.0

Comment : Polysilicon deposition

Figure 7.22.Impurity section of the ATHENA Deposit Menu.

Click on the Phosphorus checkbox and set the doping level (e.g., 5.0x1019) using the slider and the Exp menu. You can set a non-uniform grid in the deposited layer by changing the Nominal grid spacing(DY) and the Grid spacing location(YDY) parameters. To create a finer grid at the polysilicon surface, set the total number of grid layers to 10, the Nominal grid spacing(DY) to 0.02 µm and the Grid spacing location(YdY) to 0.0 (at the surface). Then, click on the Write button and the following deposition statement will be written in the input file as:

```
DEPOSIT POLY THICK=0.5 C.PHOSPHOR=5.0E19
DIVISIONS=10 \ DY=0.02 YDY=0.0
MIN.SPACING=0.001
```

Use the **Cont** button to continue the ATHENA simulation. This will create the three layer structure shown in the left plot of Figure 7.23 (Topmost layer is PolySi=0.5μm, Gate Oxide=0.02μm, Silicon = 1μm). The MIN.SPACING parameter preserves the horizontal mesh spacing for high aspect ratio grids. ATHENA tries to reduce high aspect ratio grids and MIN.SPACING stops this. To get a finer grid not at the polysilicon surface but in the middle of polysilicon layer, change YDY to 0.2. This puts on a finer grid at a distance of 0.2μm from the surface of the structure. You can do this by positioning the cursor in the input file and backspacing over existing text, or entering new text. For example:

```
DEPOSIT POLY THICK=0.5 C.PHOSPHOR=5.0E19
DIVISIONS=10 \ DY=0.02 YDY=0.2
```

It is possible to see the effect of changing the YDY. parameter within the polysilicon without rerunning the whole input file. To do this, highlight the previous statement (DEPOSIT OXIDE...), select Main Control→Init from History button, and press the

Cont button. The new history file can then be loaded into TONYPLOT (see the right plot in Figure 7.23).

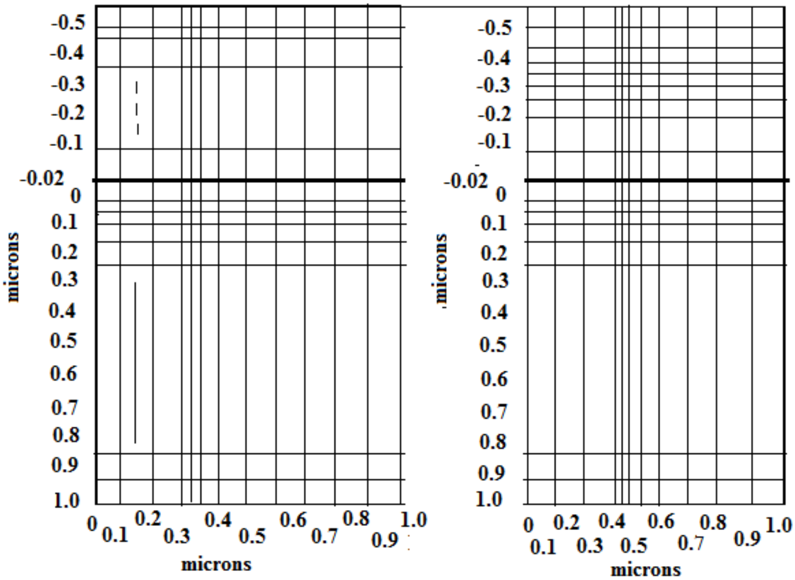


Figure 7.23 Data from History.14

Data from History.13

SSPD_Chapter 6_Part 7_Introduction to ATHENA 2_Process Simulator

SSPD_Chapter 6_Part 7 is continuation of "Creating Device Structures".Here we give 'geometrical etch','reducing grid points','reflecting the half structure to get the full structure','specifying the electrodes' and 'saving a structure file'.

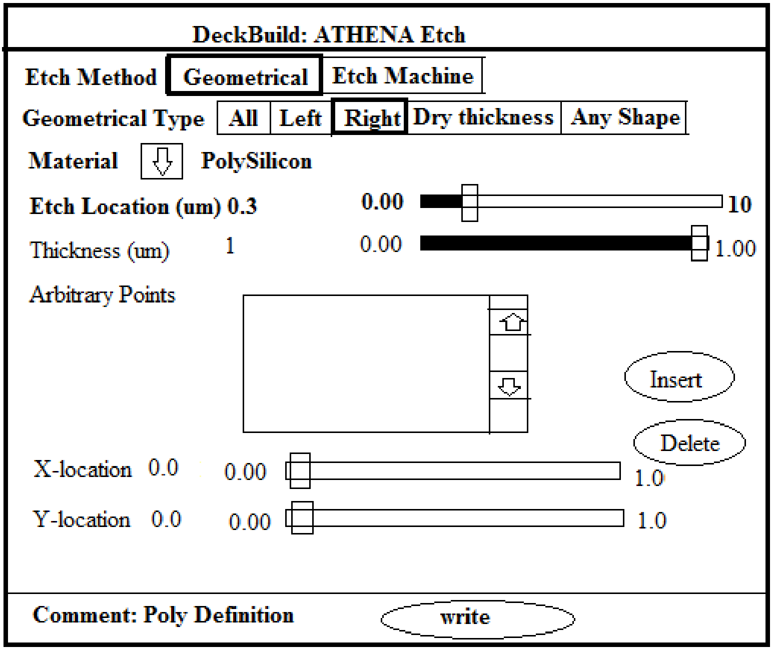
SSPD_Chapter 6_Part 7_Introduction to ATHENA2

SECTION 7.7 Background of ATHENAcontinued.

7.7.2. Creating Device Structure using ATHENA

7.7.2.5. Simple Geometrical Etches

The next step in this tutorial process simulation is to define the polysilicon gate definition. (Implant and thermal steps will be discussed in later Section :“Choosing Models In SSUPREM4”). To set a geometrical etch step, select Process→Etch→Etch... from the Command menu of DECKBUILD. The ATHENA Etch Menu (Figure 7.24) will appear.



Material PolySilicon

Etch Location (um) 0.3

Thickness (um) 1

X-location 0.0

Y-location 0.0

Comment: Poly Definition

Figure 7.24. ATHENA Etch Menu

The Geometrical etch is the default method. Other methods will be discussed in later Section called "Deposition and Wet/Dry Etching using the Physical Models in ATHENA/ELITE". Select Polysilicon from the Material menu. This example will use a polysilicon gate edge at x=0.3 and set the center of the gate at x=0.0 for the initial grid. Therefore, polysilicon should be etched to the right from x=0.3. To do so, select **Right** from the **Geometrical type**, and set the **Etch location** to 0.3. This will give the following statement: # POLY DEFINITION

ETCH POLY RIGHT P1.X=0.3

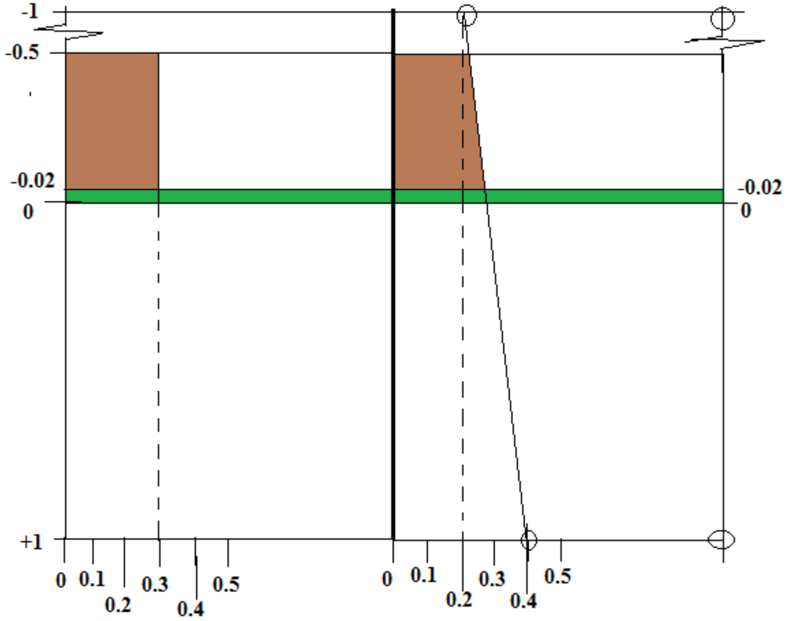
The structure created by this ETCH statement is shown in the left hand plot of Figure 7.25.

You can obtain an arbitrary shape of geometrical etching by using the **Any Shape** button. For example, to make a tilted etch, specify X and Y locations of four Arbitrary points as shown in Figure 7.26.

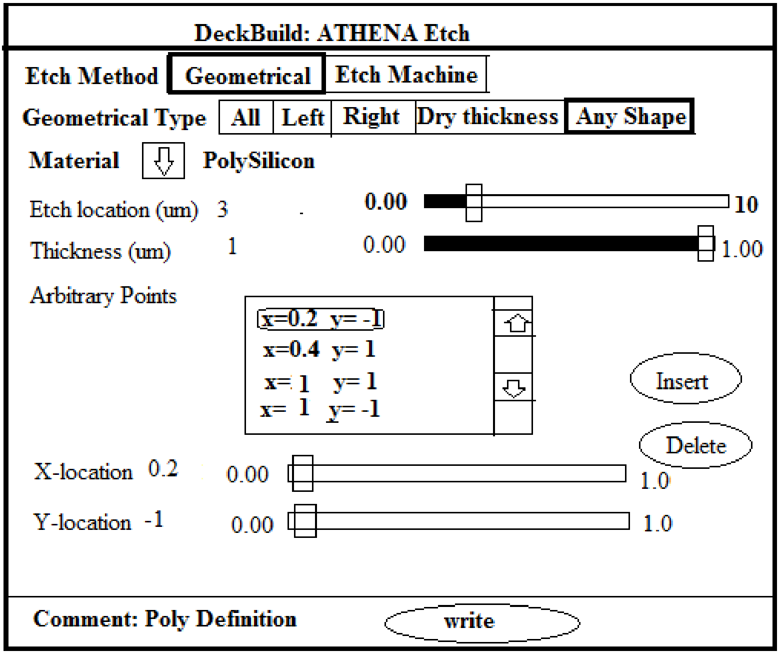
The following four etch lines will be inserted into the input file:

```
# POLY DEFINITION

ETCH POLY START X=0.2 Y= -1

ETCH CONT X=0.4 Y= 1

ETCH CONT X=1 Y= 1

ETCH DONE X=1 Y= -1
```

If this input file fragment is run instead of the previous one (using the INIT statement from the History capability), the structure after this etch sequence will appear as displayed in the right hand plot in Figure 7.25. ATHENA etches all polysilicon material within the specified polygon [polygon is defined by the points P1(x= 0.2,y= -1),P2(0.4, +1),P3(1,1) & P4(1,-1)]. The polygon etch can consist of any number of points. If you use the Insert button, an additional point will appear after the currently selected point.



Figur7.25: Structure Created by Etch Statement (a)Right etch (b) Any Shape.



Etch Method: Geometrical

Geometrical Type: Any Shape

Material: PolySilicon

Etch location (um) 3

Thickness (um) 1

Arbitrary Points

x=0.2 y= -1
x=0.4 y= 1
x= 1 y= 1
x= 1 y= -1

X-location 0.2

Y-location -1

Comment: Poly Definition

Figure 7.26. ATHENA Etch Menu for arbitrary etch.

An additional option for geometrical etching is a dry etch with a specified thickness. This can be used for spacer formation as follows: deposit an oxide of a specified thickness (e.g., 0.5μm) right to poly-Si and then etch the same thickness again beyond x = 0.6μm.

```
# CLEAN GATE OXIDE

ETCH OXIDE DRY THICK=0.02

# SPACER DEPOSITION
```

```
DEPOSIT OXIDE THICK=0.5 DIVISIONS=8

# SPACER ETCHING

ETCH OXIDE DRY THICK=0.53
```

The dry etching step etches the specified material in the region between the top (exposed) boundary of the structure and a line obtained by translating the boundary line down in the Y direction. The etch distance is specified by the THICK parameter. Figure 7-27 shows the resulting spacer.Figure 7.27 is given in the next module.

SSPD_Chapter 6_Part 7_Introduction to ATHENA3_Process Simulator
SSPD_Chapter 6_Part 7_Introduction to ATHENA3 deals with reducing the grid points

SSPD_Chapter 6_Part 7_Introduction to ATHENA2

SECTION 7.7 Background of ATHENAcontinued.

7.7.2. Creating Device Structure using ATHENA

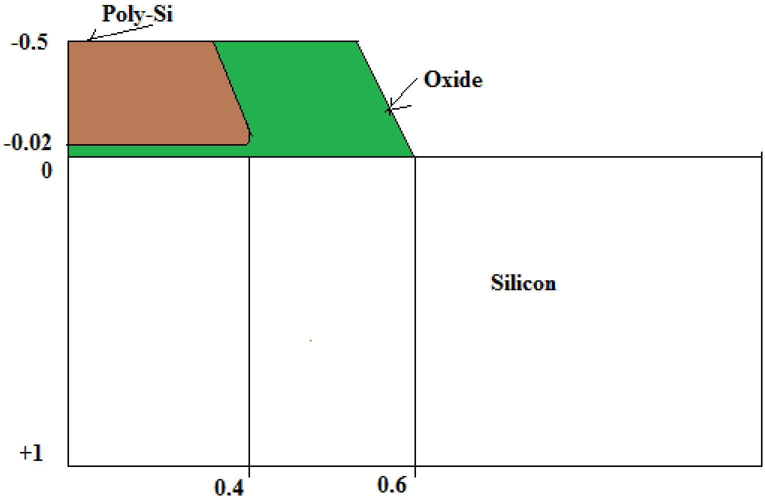


Figure 7.27. Spacer Formation using dry etch.

7.7.2.6. Reducing Grid Points in Non-Essential Areas using the Relax Parameter

The previous sections demonstrate that the quality of the grid is extremely important for ATHENA simulation. The rectangular based grid generated by the INITIALIZE or DEPOSIT statements may remain intact in those areas not involved in the process steps affecting the grid (e.g., etching or oxidation). The Grid Relax capability allows the spacing to be increased in such areas at any point which are not critical in the process step during the simulation. This capability is useful for two reasons. First, the initial small spacings are propagated throughout the structure. For example, the fine grid in the X-direction shown in Figure 7.23 may be needed only in the upper portion of the structure where doping occurs. Eliminating some grid lines and points in the lower portion of the structure will not affect the accuracy of implant and diffusion simulation. Second, it is always necessary to set a fine grid in the area where ion implantation takes place, but the fine grid may be unnecessary after the profile is levelled-off during thermal steps. So, relaxation of an initially fine grid may save simulation time during subsequent steps. Parameters for the RELAX statement are set from the ATHENA Relax Menu.

To open this menu, select Structure → Relax... in the DECKBUILD Commands menu. You can preform grid relaxation over the whole structure if you select Entire Grid, or within a selected rectangular area, if you choose Selected and specify Xmin, Xmax, Ymin, and Ymax in Location Select. Selecting a material from the Material menu specifies which material region will be affected by the grid relax operation.

The default is all materials within the specified area. You can perform the grid line elimination either in one direction or in both directions by selecting X direction, Y direction, or Both. The X direction, Relax, cannot be performed for individual materials except for the substrate.

To understand how the relax function changes a grid, we will use the structure we have obtained after spacer formation was specified . If we relax the entire grid in both directions , the following lines will be inserted into the tutorial input file:

```
# RELAX EVERYWHERE RELAX DIR.X=T DIR.Y=T
```

Here ‘T’ is True and ‘F’ is False.

The total number of grid points is reduced from 708 to 388. When comparing with the grid before relaxation, note that the grid within the oxide spacer and polygate has not changed. This is due to three factors:

- the relax algorithm works only with rectangular base grid,

- it never eliminates grid lines adjacent to a region boundary,

- the relaxed area should be at least five by five grid points.

Within silicon, each second horizontal line is eliminated. The lower part of each second vertical line is also eliminated. This happens because the algorithm doesn’t allow the formation of obtuse triangles.

If you don’t want the grid to be relaxed above y = 0.3, choose Selected and set all four boundaries in the Relax box. This will give the following RELAX statement:

```
# RELAX LOWER HALF OF THE STRUCTURE

RELAX X.MIN=0.00 X.MAX=1.00 Y.MIN=0.3
Y.MAX=1.00 DIR.X=T DIR.Y=T
```

In this case, the number of grid points is 567. The grid above y = 0.3 remains intact and the elimination in X and Y- directions happens only below y = 0.3.

To increase spacing only in the X-direction in the area below y = 0.3, select the X direction and leave the Area and Location Selections as before. This will give the following Relax statement:

```
# RELAX LOWER HALF ONLY IN X-DIRECTION

RELAX X.MIN=0.00 X.MAX=1.00 Y.MIN=0.3
Y.MAX=1.00 DIR.X=T DIR.Y=F
```

Note: The only difference is that instead of

DIR.Y=T, the statement contains DIR.Y=F, which prevents elimination in Y- direction. This gives 638 grid points and a different pattern of elimination

You can also apply several consequent RELAX statements to achieve grid elimination in different areas of the structure.

An important thing to remember about the RELAX capability is that it allows you to avoid creating obtuse triangles and avoid relaxing directly on the material boundaries. This sometimes results in no relaxation or grid relaxation in a subset of the desired area. The most desirable method for complete control over gridding is by using DEVEDIT as described briefly in this chapter and in the VWF INTERACTIVE TOOLS User's Manual, Vol. II.

SSPD_Chapter 6_Part 7_Introduction to ATHENA 4_Process Simulator

SSPD_Chapter 6_Part 7_Introduction to ATHENA 4 deals with reflecting the half MOS structure to obtain the full structure through mirror parameter.

SSPD_Chapter 6_Part 7_Introduction to ATHENA4

SECTION 7.7 Background of ATHENAcontinued.

7.7.2. Creating Device Structure using ATHENA

7.7.2.7. Reflecting a Structure in the “Y” Plane using the Mirror Parameter

This tutorial process simulation has been building one half of a MOSFET- like structure. At some point in the simulation, you will need to obtain the full structure. This must be done before exporting the structure to a device simulator or setting electrode names. In general, structure reflection should be performed when the structure ceases to be symmetrical (e.g., a tilted implant, an asymmetrical etching, or a deposition takes place), or when a reflecting boundary condition no longer applies to the side, which is going to be the center of the structure.

This example will explain how to mirror the structure at its left boundary. To mirror the structure, select Structure → Mirror in the

Commands menu (Figure 7.28).

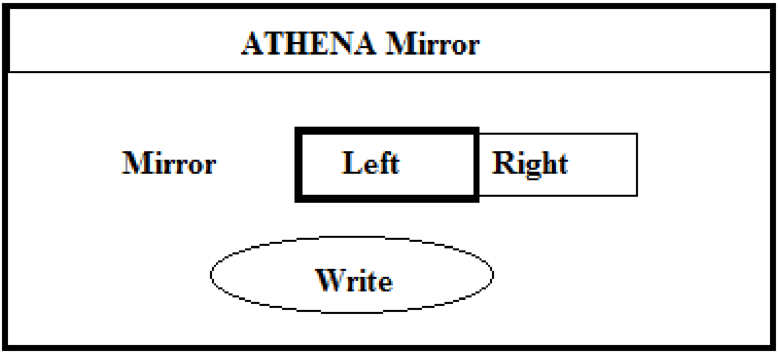


Figure 7.28. ATHENA Mirror Menu

Then, press the Write button to write the following statement to the input file:

```
STRUCT MIRROR LEFT
```

The resulting structure is shown in Figure 7.29.

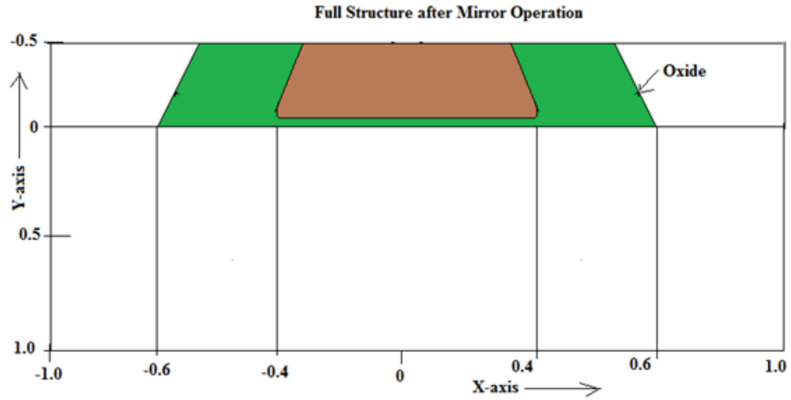


Figure 7.29. ATHENA Reflect capability

The left half of the structure is a complete mirror copy of the right part, including node coordinates, doping values, and so on. Beware of rounding errors when mirroring. If the boundary of reflection is not smooth to within 0.1 angstroms, some points will be duplicated.

SSPD_Chapter 6_Part 7_Introduction to ATHENA 5_Process Simulator

SSPD_Chapter 6_Part 7_Introduction to ATHENA 5 deals with specification of electrodes and saving the structure file.

SSPD_Chapter 6_Part 7_Introduction to ATHENA2

SECTION 7.7 Background of ATHENAcontinued.

7.7.2. Creating Device Structure using ATHENA

7.7.2.8 Specification of Electrodes in ATHENA

The ultimate goal of an ATHENA simulation is usually to create a device structure (material layers plus doping), which then can be used by a device simulator (usually ATLAS) for electrical characterization. Although ATLAS is able to specify the locations of electrodes, in many cases specifying electrodes must be done in ATHENA. For example, it is impossible to specify an electrode location in ATLAS when the electrode does not consist of straight segments. Also, when specifying electrodes in ATHENA, it is useful to transfer electrode layer information from layout to electrical tests in a device simulator

ATHENA can attribute an electrode to any metal, silicide, or polysilicon region. A special case is the backside electrode, which can be placed at the bottom of the structure without having a metal region there. If you deposit 0.1 μm aluminum layer on the full structure after reflection using:

```
DEPOSIT ALUMIN THICK = 0.1
```

and etched the following part of the layer between x= -0.8 and x= 0.8, using the Any Shape specification in the Athena Etch Menu (See Figure 7.26):

```
ETCH ALUMINUM START X= -0.8 Y= -1.0

ETCH CONT X= -0.8 Y= 1.0

ETCH CONT X= 0.8 Y= 1.0

ETCH DONE X= 0.8 Y= -1.0
```

you will now have the structure shown in Figure 7.30

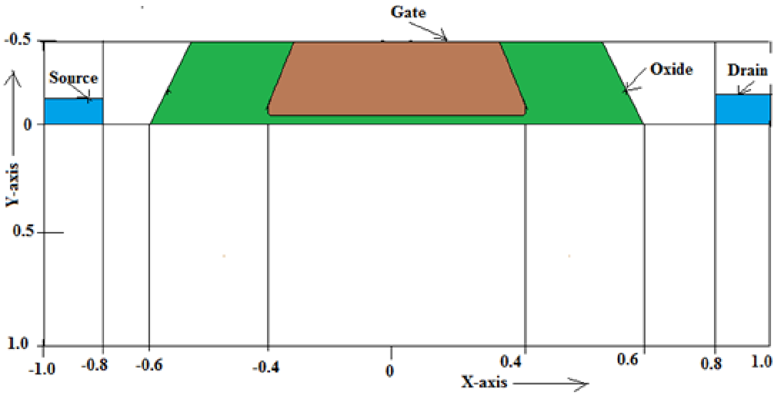
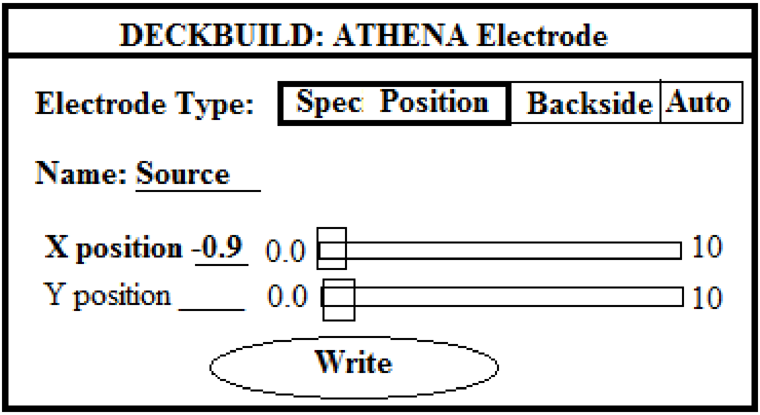


Figure 7.30. MOSFET structure with Electrodes.

You can now use the ATHENA Electrode menu (see Figure 7.31) by selecting Commands→Structure→ Electrode....



Electrode Type: Spec Position

Name: Source

X position -0.9

Figure 7.31.ATHENA Electrode Menu

To set an electrode at a specified position, select the Specified Position button, type in the X Position (e.g., -0.9) and Name (for example, source), and press the Write button. The following statement will appear in the input file:

ELECTRODE NAME=SOURCE X= -0.9

Similarly, specify the drain electrode:

ELECTRODE NAME=DRAIN X= 0.9

The polysilicon gate electrode specification has the same format. For this structure it can be done the same way as for source or drain:

```
ELECTRODE NAME= GATE X= 0.0
```

If the polysilicon layer is not the topmost layer at x=0, the Y Position can be specified. In this case, check the Y Position checkbox and type in a y coordinate within the polygate layer (e.g., -0.2). If Y is not specified and the electrode is not on top, ATHENA will look for the electrode in the underlying layers. If it fails, an error will be reported.

To specify a backside electrode, select Backside from the Electrode Type field and type in a name (see Figure 7.31) .

The following backside electrode statement will appear in the input file:

```
ELECTRODE NAME=BACK BACKSIDE
```

If an electrode name is not specified, DECKBUILD issues the error message:

NO ELECTRODE NAME SPECIFIED and the command is not written to the input file.

If an incorrect position for electrode is specified, for example:

```
ELECTRODE NAME=JUNK X=0.6
```

ATHENA will output the following warning message: Cannot find the electrode for this structure. Electrode statement ignored and ignores the statement.

7.7.2.9. Saving a Structure File for Plotting or Initializing an ATHENA Input file for Further Processing

As mentioned in the “Standard Structure File Format” Section, the DECKBUILD history function saves structure files after each process step. In many cases, however, you need to save and initialize structures independently. There are several reason why it’s needed to save and initialize structures independently.

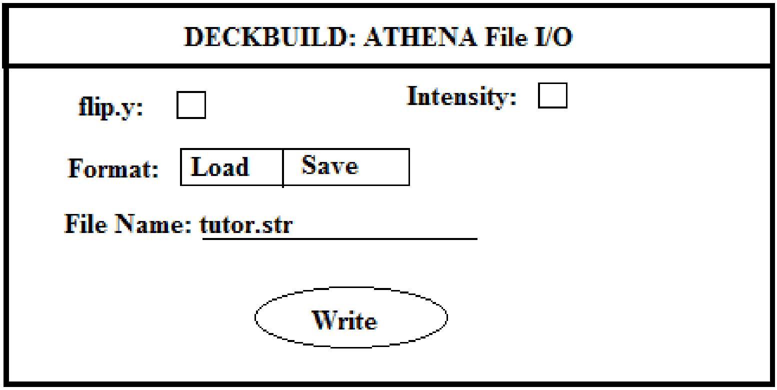
The first reason is because the stack for the history files is limited (50 by default). The second reason is because it is usually undesirable to keep dozens of history files on disc (each of which occupy hundreds of Kbytes) after the DECKBUILD session ends. The third reason is because users often want to save the structure information generated after key process steps (e.g., final structure)

To save or load a structure, use the ATHENA File I/O Menu (See Figure 7.32) by selecting Commands → File I/O.... Specify a file name (the file extension.str is recommended for all ATHENA structure files) and press the Save button. The following line will appear in the input file:

```
STRUCT OUTFILE=TUTOR.STR
```

You can reload this file (tutor.str) back into ATHENA at any time during the current DECKBUILD session or during any subsequent session. To reload the structure file, press the Load button on the ATHENA File I/O menu. The following INIT statement will appear:

```
INIT INFILE=TUTOR.STR
```



File Name: tutor.str

Figure 7.32. ATHENA File I/O Menu

Note: Only the structure will be reloaded if ATHENA is restarted before this INIT statement. Any parameters or coefficients that were set during previous simulations must be reset if they are needed. This structure file can also be used by any device simulator or DEVEDIT.

SSPD_Chapter 6_Part 7_Introduction to ATHENA 6_Process Simulator

SSPD_Chapter 6_Part 7_Introduction to ATHENA6_Process Simulator describes the need for multiple models for each process, choosing appropriate model with method statement,method statements during process flow,switching guidelines, modeling the correct substrate depth, simulating ion implantation, simulating ion implantation damage, simulating diffusion, simulating rapid thermal anneal and simulating oxidation.

SSPD_Chapter 6_Part 7_Introduction to ATHENA 6_Process Simulator.

7.7. Background of ATHENA.

7.7.3. Choosing Models In SSUPREM4

This section describes how to simulate process steps (e.g., implantation, diffusion/oxidation, epitaxy, and silicidation) specific to the SSUPREM4 module of ATHENA. Also discussed, are the SSUPREM4 model statements, such as METHOD, OXIDE, MATERIAL, and IMPURITY.

When simulating any process involving dopant or its diffusion or both, it is absolutely critical for simulation accuracy to use the appropriate model. Process steps where correct choice of models are vital include implantation, diffusion, rapid thermal annealing, oxidation, and epitaxy . This section provides specific advice on which models should be used for each process step.

7.7.3.1: The Reason for Multiple Models for Each Process

The key to simulating any dopant related process is to accurately account for damage in the semiconductor. For example, in silicon processing, typical implantation doses can cause sufficient damage to the substrate to enhance dopant diffusion rates by three orders of magnitude or more, so choosing the wrong model in this instance will result in inaccurate results.

Well known device anomalies such as the Reverse Short Channel Effect in MOS processing or the emitter push effect in bipolar processing are wholly the result of such damage enhanced diffusion.

Other processes that consume the semiconductor, such as oxidation and silicidation also inject damage into the substrate. This must be accounted for if accurate dopant profiles are a requirement.

This section aims to provide you with a set of rules outlined, indicating the correct model that can be used most of the time without you having a detailed knowledge of the physics involved. The usual rules of model selection apply here. The more complicated the model, the greater the simulation time. There is always a compromise between simulation accuracy and simulation time. The following sections describe when to use the hierarchy of models so that the most complicated models are only used when you make a significant difference to the result.

7.7.3.2.: Choosing an Appropriate Model Using the Method Statement

The hierarchy of diffusion and damage models available is broadly related to the maximum level of damage already in the semiconductor or the maximum level of damage that the next process step is likely to introduce at any particular time during the process flow. The level of damage in the semiconductor at any one time is not a static quantity but will depend on when and how much damage was induced by a process step and how much annealing has occurred in subsequent thermal steps. The range of models available to you can account for all of the above effects and allows accurate simulation of dopant diffusion if appropriate models have been chosen.

The choice of model or combination of models for any of the process steps described above is defined in the METHOD statement. The METHOD statement serves a number of functions but in the context of defining damage models the METHOD statement is used for two purposes.

The first purpose is to specify models for how damage is induced during processes such as implantation or oxidation. The second purpose is to specify how that damage anneals and diffuses in subsequent or concurrent thermal processes.

It's important to realize that the METHOD statement must be placed above the line, specifying the process step or steps to which it refers in the input file. Any number of method statements can be used in an input file allowing you to change the models at will during the process flow to optimize the speed and accuracy of the simulation. The models specified in the METHOD statement will hold true for all processes that follow it until it's updated by a subsequent method statement.

Table 7.5 below indicates a recommended method statement for typical processes. It should be realized that these statements are hierarchical, so there is no accuracy lost if a more complicated model is used where a simpler one would suffice. The only downside here is a longer simulation time. The table below starts off with the simplest of models and progresses to the more complicated ones.

Table 7.5. Recommended Method Statements for Typical Processes.

Method Statement Syntax	Suitability of using this method
method fermi	Use only before undamaged silicon diffusions, where doping concentrations are less than 1e20/cm3 and no oxidizing ambient is present.
method two.dim	Use before implant doses less than 1e13/cm2 and for oxidations.
method full.cpl cluster.dam high.conc	Use before implant doses greater than 1e13/cm2

7.7.3.3. : Changing the Method Statement During the Process Flow

It has previously been stated that the disadvantage of using the most advanced and complex models is the time involved during diffusion cycle simulation. Accordingly, there is an incentive during complex process simulations to switch back to a simpler model during a diffusion cycle when the majority of the damage created by a previous implant has been annealed. We will show you when to switch to a simpler model.

If the process being modeled has involved implantation or oxidation at any stage, we advise not to use the fermi model. An exception to this would be in some power devices with very long diffusion times where the exact nature of surface damage would have little impact on the final distribution of the dopant and simulation time is at a premium.

In reality, for most small geometry processes, the question of switching models becomes one of when to add a new method statement that changes from:

```
METHOD FULL.CPL CLUSTER.DAM HIGH.CONC
```

to

```
METHOD TWO.DIM
```

after a high dose implant.

7.7.3.4. Switching Guidelines.

A simple guideline to follow when to switch method statements during a process flow is by switching back to the TWO.DIM model if the anneal temperature is greater than 900° and the device has been annealed for at least one minute, following an implant where the dose is greater than 1e13/cm^2.

For a more accurate guideline, see the Chapter 3: “RTA Diffusion Modelling”. Table 3-6 shows the anneal temperature/time combinations required for 95% of the clusters formed during high dose implants to dissolve. Modeling these dopant/defect clusters requires the fully coupled (full.cpl) and cluster damage (cluster.dam) models. Only when these clusters have dissolved can the two.dim model be used without significant loss of simulation accuracy. As a general rule, we recommend that the method statement be changed to method two.dim only after a diffusion time that is at least two or three times as long as the values quoted in the table.

If you wish to be certain of when it's safe to switch models, the recommended procedure is to save a structure file at the point of interest, load the file into TONYPLOT and perform a 1D cutline. Plot the clusters and interstitials. If the cluster concentration is still visible, it's too early to switch models.

For power devices, where simulation time is at a premium, the same method already described should be used. But instead of using the cluster concentration as a guide of when to switch models, the interstitial concentration should be used as the guide as to when to switch models one more time from the TWO.DIM model to the basic FERMI model. When the interstitial concentration near the surface during a very long anneal has been reduced to only marginally above the background level at the anneal temperature concerned, the method statement can be switched to METHOD FERMI to greatly reduce the simulation time. The interstitial background level will be the level deep in the substrate where little damage has occurred.

7.7.4: Modelling the Correct Substrate Depth

An important and often overlooked aspect of the correct modeling of dopant diffusion is the choice of substrate depth. It has been mentioned previously that the rate of dopant diffusion is highly dependent on the level of damage in the substrate. Therefore, the accurate modeling of dopant diffusion requires the accurate modeling of substrate damage, particularly the movement of interstitials. In general, the interstitials created directly or indirectly by implantation and oxidation tend to diffuse much greater distances than the dopant. The substrate depth chosen for modeling purposes must therefore be deep enough to allow the interstitial concentrations to return to background levels at the bottom of the simulated substrate, even if no dopant diffusion occurs at this depth.

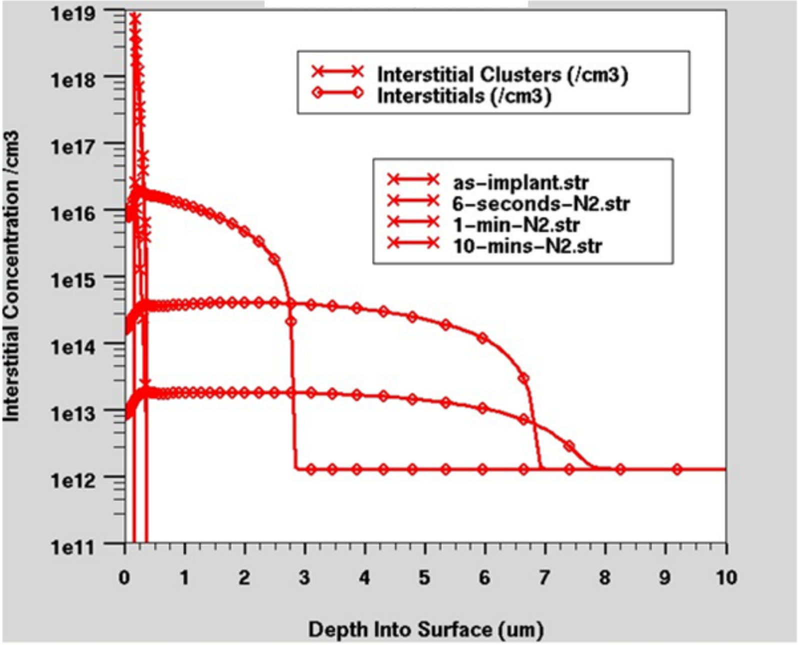


Figure 7.33. Interstitials can move far into the substrate even after a short 10min anneal

Figure 7.33 shows typical diffusion profiles of interstitials after a 1e15/cm3 20keV Boron implant at various anneal times (as-implant, 6 seconds anneal, 1 minute anneal and 10 minute anneal). After only a 10 minute anneal, the interstitials have diffused 8µm into the substrate.

Interstitials, like dopant, require a concentration gradient in order for overall diffusion to take place. For example, if the concentration gradient of interstitials is removed by having too shallow a substrate depth, the concentration of interstitials will start to pile up because they are no longer being removed through diffusion into the bulk of the substrate. If the level of modeled interstitials becomes too high, the diffusion of dopant, even near the surface of the substrate, will also be too high and the simulation will be inaccurate as shown in Figure 7.34.

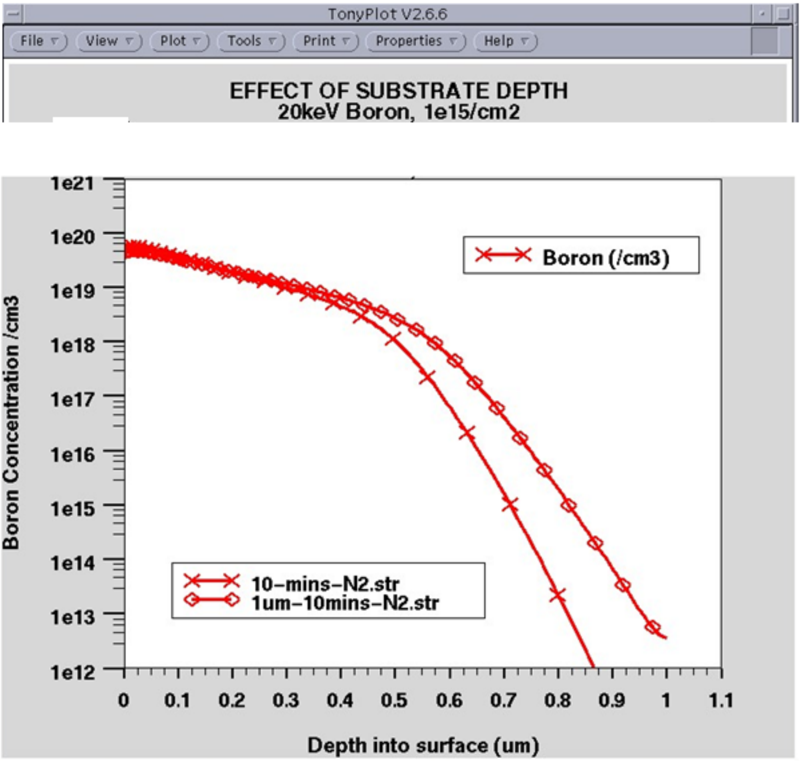


Figure 7.34. Effect on Boron diffusion profile when too small a substrate depth is used in the simulation

Figure 7.34. shows the Boron profiles for two identical anneals, the only difference is the depth of the simulated substrate. You’ll see that a shallow modeled substrate (the upper curve with circles) always results in more total diffusion, even though the substrate depth was greater than the total diffusion depth in both cases.

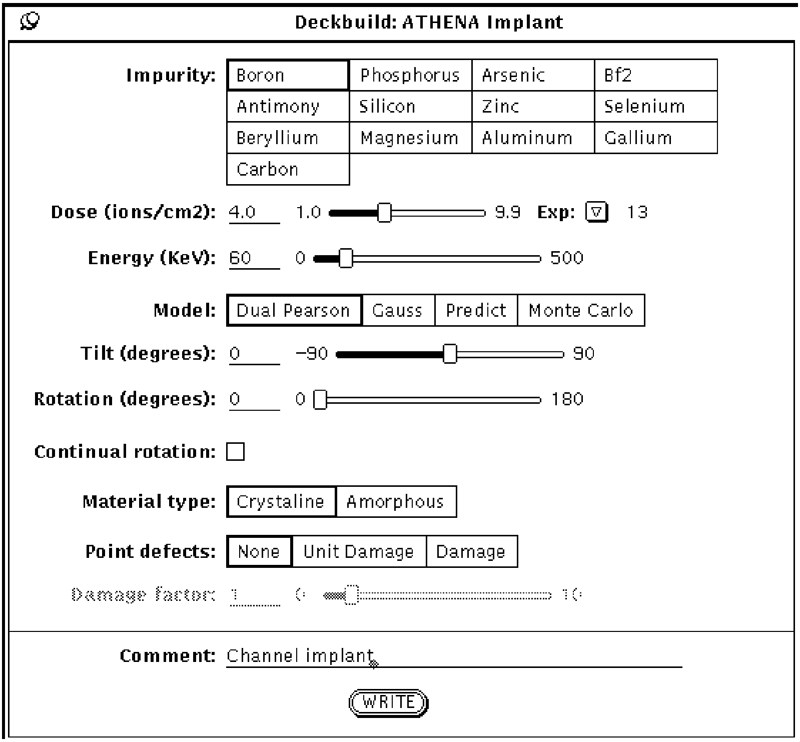
Modeling a deep substrate doesn’t need to involve a huge number of extra mesh points, since the mesh points can be placed quite far apart near the bottom of the substrate. All that is required of the mesh points near the bottom of the substrate is that there be sufficient to model the gradient of interstitials in this region. The number of additional mesh points can be further reduced in the X-direction by the using several RELAX statements.

For normal small geometry MOSFET/Bipolar processing, a substrate depth of 20 µm should be more than adequate. This depth can be reduced by plotting the vertical interstitial profiles at various points in the process to find the maximum depth of interstitial diffusion. There is little to be gained by reducing the depth of simulation, however, if the combination of large grid spacing is deep in the substrate and the RELAX statement is used appropriately.

7.7.5. Simulating Ion Implantation

Ion implantation is the main method used to introduce doping impurities into semiconductor device structures. Adequate simulation of the ion implantation process is very important because modern technologies employ small critical dimensions (CDs) and shallow doping profiles, high doses, tilted implants and other advanced methods.

The IMPLANT statement can be set by using the ATHENA Implant Menu (Figure 7.35). To open this menu, select Process →Implant... in the Commands menu.



Impurity: Boron

Dose (ions/cm2): 4.0 Exp: 13

Energy (KeV): 60

Model: Dual Pearson

Tilt (degrees): 0

Rotation (degrees): 0

Continual rotation: ☐

Material type: Crystaline

Point defects: None

Damage factor: 1

Comment: Channel implant

Figure 7.35. ATHENA Implant Menu.

The following gives the minimum set of parameters that should be specified:

- Name of implant impurity (e.g., boron)

- Implant dose using the slider for the pre-exponential value (e.g., 4.0) and the Exp menu for the exponent (e.g., 12)

- Implant energy in KeV (e.g., 60)

- Tilt angle in degrees (e.g., 7°)

- Rotation angle in degrees (e.g., 30°)

All other parameters can use their default values.

Press the Write button and the following statement will appear in the input file.

```
# CHANNEL IMPLANT

# DUAL PEARSON MODEL IS BEING USED FOR
IMPLANT PROFILE

IMPLANT BORON DOSE= 4.0E12 ENERGY= 60
PEARSON TILT=7 ROTATION=30 \ CRYSTAL
```

All of the parameters in the statement above are self-explanatory except CRYSTALLINE. The CRYSTALLINE parameter indicates that for all analytical models, the range statistics extracted for a single silicon crystal will be applied (when available). If AMORPHOUS is selected, the range parameters measured in pre-amorphized silicon will be used (when available). The CRYSTALLINE parameter also has another meaning for the Monte Carlo or BCA implant models. It invokes the Crystalline Material Model which takes channeling into account. Note that the latter model is much slower (5 - 10 times) than the Amorphous Material Model. The Crystalline Material Model is the default model for BCA or Monte Carlo simulation.

For a detailed description of ion implant model selection, see Chapter 3: “SSUPREM4 Models”, Section 3.5:“Ion Implantation Models”.

You can specify tilt and rotation angles of the ion beam. Positive tilt angles correspond to the ion beam coming from the top left. Specifying the rotation angle makes sense only for non-zero tilt angles. Zero rotation means that the ion beam vector lies in the plane parallel to the 2D simulation plane. 90° rotation means that the ion beam vector lies in the plane perpendicular to the simulation plane.

Selecting Continual rotation causes SSUPREM4 to rotate the wafer, i.e., implantation will be performed at 24 different rotation angles from 0 to 345°, in increments of 15°.

There are several damage models available in

SSUPREM4. These models allow you to estimate distributions of various defects generated after ion implantation. For more details about the damage models and their effect on subsequent diffusion, see Chapter 3: “SSUPREM4 Models”, Section

7.7.6:“Ion Implantation Damage”.

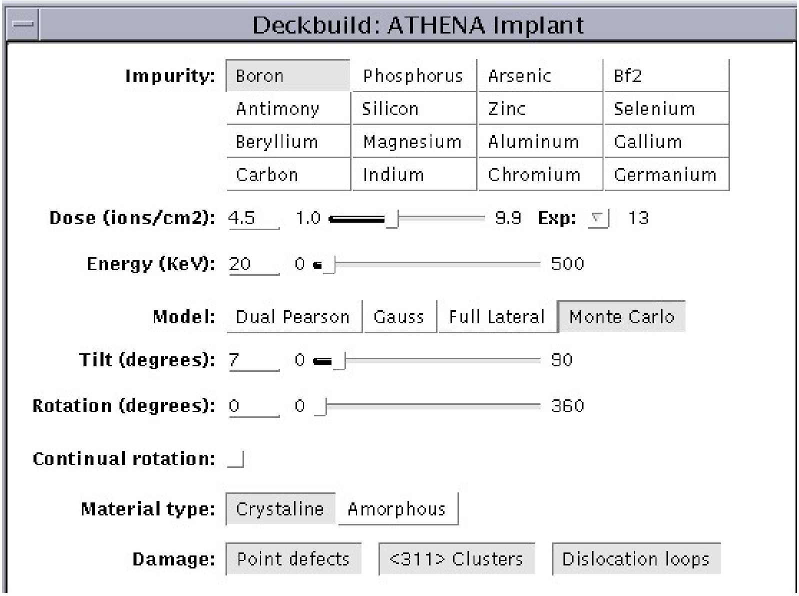
When the Monte Carlo model is selected, you can specify several additional optional parameters (See Figure 7.36). The first three parameters are related to the Damage model (Point defects, {311}-clusters, and dislocation loops). The three others control Monte Carlo calculation (initial random number, number of trajectories, and smoothing). See Table 7.6 for a quick reference of ATHENA implant models.

Table 7.6. ATHENA Implant Model Reference

Process	Model	Assumption	Recommendation
Implant	SIMS Verified DualPearson (SVDP) - Default	Empirical	See Chapter 3: “SSUPREM4 Models”, Table 3.7.
	Single	Analytic	All other

	Pearson Monte CarloMonte or BCA	Statistical	cases Multi-layer structures: angled implants into a structure where many ions could be reflected (trenches); when channeling is not described by SVDP; high or very low energy
SiliconType	Amorphous	No channeling effect is included	Most of implant profile is within amorphous materials (oxide, polysilicon, pre-amorphized silicon); channeling is negligible

			or not important
	Crystal - Default	Channeling effect is included	When channeling effects are important: light ions (boron, phosphorus)_, zero or close to 0° tilt, implant through thin amorphous layer into crystalline substrate



Impurity: Boron

Dose (ions/cm2): 4.5 Exp: 13

Energy (KeV): 20

Model: Monte Carlo

Tilt (degrees): 7

Rotation (degrees): 0

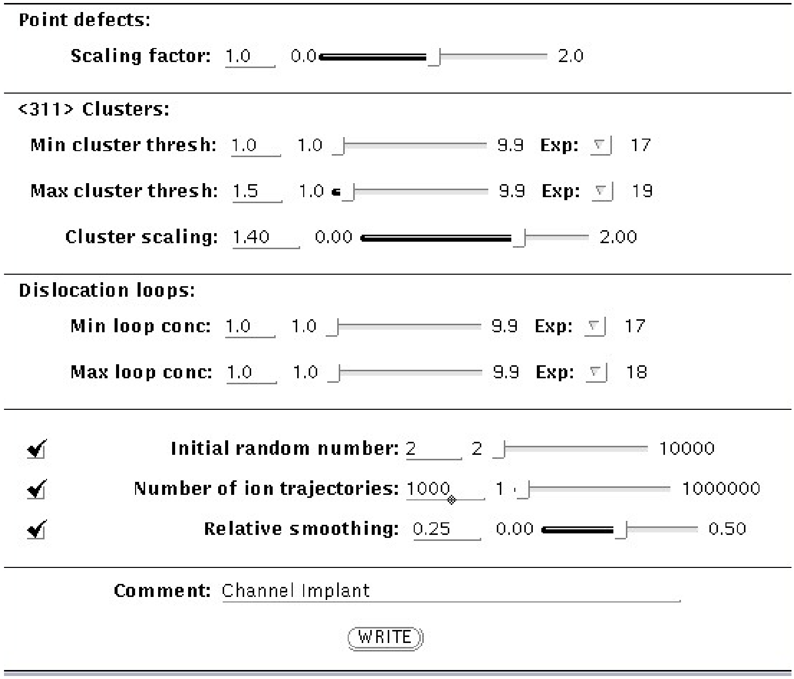
Continual rotation: ☐

Material type: Crystaline

Damage: Point defects <311> Clusters Dislocation loops

Figure 7.36. ATHENA Implant Window

continued on next page



Point defects:

Scaling factor: 1.0 0.0 2.0

<311> Clusters:

Min cluster thresh: 1.0 1.0 9.9 **Exp:** 17

Max cluster thresh: 1.5 1.0 9.9 **Exp:** 19

Cluster scaling: 1.40 0.00 2.00

Dislocation loops:

Min loop conc: 1.0 1.0 9.9 **Exp:** 17

Max loop conc: 1.0 1.0 9.9 **Exp:** 18

☑ **Initial random number:** 2 2 10000

☑ **Number of ion trajectories:** 1000 1 1000000

☑ **Relative smoothing:** 0.25 0.00 0.50

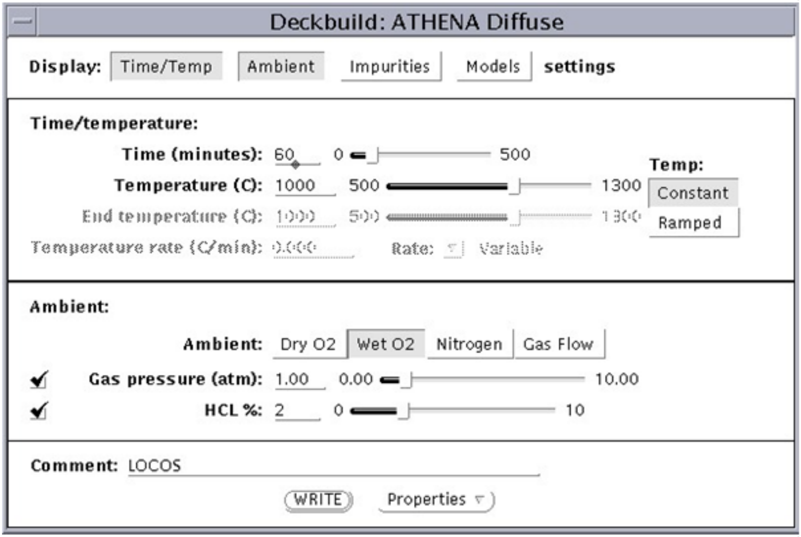
Comment: Channel Implant

Figure 7.36. ATHENA Implant Window

7.7.7 Simulating Diffusion

Simulation of thermal process steps is a focal point of SSUPREM4. The hierarchy of diffusion and oxidation models is described in this chapter and in Chapter 3: “SSUPREM4 Models”, Sections 3.1: “Diffusion Models” and 3.3: “Oxidation Models”. This section will demonstrate how to set different parameters and models of diffusion, oxidation and silicidation. The last process will take place only if at least one refractory metal or silicide layer is present in the structure.

The parameters and models of a diffusion/oxidation step can be prepared from the ATHENA Diffuse Menu. (Figure 7.37).



Deckbuild: ATHENA Diffuse

Display: Time/Temp Ambient Impurities Models settings

Time/temperature:

Time (minutes): 60 0 500

Temperature (C): 1000 500 1300

End temperature (C): 1000 500 1300

Temperature rate (C/min): 0.000 Rate: Variable

Temp: Constant Ramped

Ambient:

Ambient: Dry O2 Wet O2 Nitrogen Gas Flow

✓ Gas pressure (atm): 1.00 0.00 10.00

✓ HCL %: 2 0 10

Comment: LOCOS

Figure 7.37. ATHENA Diffuse Menu.

To open this menu, select Process→Diffuse... in the Deckbuild Commands menu. The Diffuse menu has four sections. Only the Time/Temperature and Ambient fields appear initially. The Impurities and Models fields appear only when the corresponding check boxes are selected.

The minimum set of diffusion step parameters is as follows:

- Time (e.g., 60 minutes)

- Temperature (e.g., 1100° Celsius)
- Gas pressure (1 atmosphere is default)

The following input file statements will appear:

```
# DRIVE-IN

DIFFUSE TIME=60 TEMP=1100 NITRO
PRESS=1.00
```

If you choose the Ramped box and End Temperature or Temperature rate, a ramped temperature thermal step is simulated. The temperature rate is a variable by default, but it can be set to a specific constant temperature rate by selecting Constant in the Rate box. If the End Temperature is set to 1000, the following lines appear:

```
# RAMPING DOWN

DIFFUSE TIME=60 TEMP=1100 T.FINAL=1000
NITRO PRESS=1.00
```

The same pull down menu used for inert diffusions is also used for oxidations described in the “Simulating Oxidation” Section on page 2-39. But, since there are special considerations for inert diffusions which come under the category of Rapid Thermal Anneals (RTA), the special notes pertaining to this specific set of conditions are described in the next section. These notes are very important for accurate simulation of high temperature, short duration anneals. We recommend that you read these notes before attempting to write the RTA section of the input file.

7.7.8: Simulating Rapid Thermal Anneals (RTA) Notes

The usual reason for employing a Rapid Thermal Anneal (RTA) in a process flow is to anneal out damage in the substrate that has been caused by a previous process step, usually an implant, while at the same time minimizing dopant diffusion. Dopant activation also occurs during this process. These anneals are usually high in temperature and low in duration for sound device physics reasons.

Once again, the key to accurate simulation of RTA lies in the accurate simulation of substrate damage behavior. The role of interstitials in enhanced dopant diffusion has already been explained in Section 2.4:“Choosing Models In SSUPREM4” to become familiar with the role of interstitials during process simulation.

The reason why an RTA usually employs high temperatures and short durations is because for a given high dose implant, if an anneal duration is selected so that a fixed percentage of the damage is annealed, the lower the anneal temperature, the more dopant diffusion occurs.

The above statement requires an explanation since intuitively, the opposite would seem more likely. A descriptive explanation of what is happening can be informative if the two extremes of anneal temperature are considered.

For the lowest anneal temperatures, the damage anneal rate is almost zero, so dopant diffusion rates are enhanced by a factor of 1000°C or more for the long time periods required to remove the damage. This results in high total dopant diffusion.

For the highest temperature anneals, a significant percentage of damage removal occurs in a fraction of a second. Almost zero damage enhanced diffusion or total diffusion occurs in this instance, and the anneal time to remove the damage is very short. Extrapolating between these extremes provides a qualitative explanation of what occurs for intermediate temperature anneals.

Two important points have now been established:

1. For sound device physics reasons, most RTA processes consist of high temperature, short duration anneals.

2. Damage-enhanced diffusion will only occur for a few seconds at typical RTA temperatures.

For accurate simulation of RTA, the second point is most important and often wrongly neglected.

Suppose an RTA consists of a 10 second ramp up to 1000°C, followed by a 20 second anneal and a 10 second cool down. From the second point, it is apparent that most of the Total Dopant Diffusion would have taken place during the Ramp Up Phase of the RTA.

Therefore, always model the temperature ramp up accurately when simulating an RTA process. In most cases, the ramp down can be neglected, since all the diffusion has already taken place at the beginning when the silicon was still damaged.

7.7.9: Simulating Oxidation

It has already been stated that the pull down menu for simulating oxidations is the same as that for simulating inert diffusions described in the "Simulating Diffusion" Section on page 2-37. See this section for advice on selecting the appropriate pull down menu from DECKBUILD.

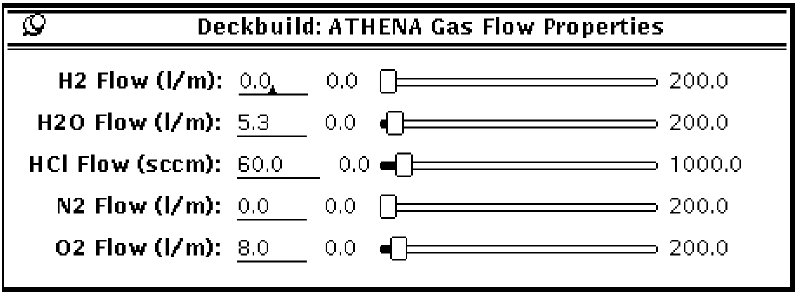
The default method for oxidation is Compress. In SSUPREM4 examples there are a number of examples which illustrate the use of different models for different processes and structures.

In our previous example described in the "Simulating Diffusion" Section on page 2-37, if the next temperature step is going to be at a constant temperature of 1000°C in dry O2 with 3% of HCL in the ambient, select the Dry O2 box and set HCL%

equal to 3 in the Ambient section of the Diffuse menu. The following input file fragment will appear:

```
# GATE OXIDE

DIFFUSE TIME=60 TEMP=1000 DRYO2
PRESS=1.00 HCL.PC=3
```

If the ambient is a mixture consisting of more than one oxidant, the total oxidation rate will depend on the combined effect of all species in the ambient. To specify the contents of the ambient mixture, select the Gas Flow button in the Ambient section and an additional ATHENA Gas Flow Properties Menu (Figure 7-38) will appear.



H2 Flow (l/m): 0.0
H2O Flow (l/m): 5.3
HCl Flow (sccm): 60.0
N2 Flow (l/m): 0.0
O2 Flow (l/m): 8.0

Figure 7.38. ATHENA Gas Flow Properties

If the Gas Flow components are selected, as shown in Figure 7.38, the following statement will be generated:

```
# GATE OXIDE
```

```
DIFFUSE TIME=60 TEMP=1000 F.H2O=5.3
F.HCL=0.06 F.O2=8.0 \ PRESS=1.00
```

One or several impurities can be present in the ambient. To set ambient in the Impurity Concentration section of the ATHENA Diffuse Menu (See Figure 7-37), check the corresponding checkboxes, and set the values using sliders and the Exp menus.

For example, by selecting the appropriate boxes and values, the following DIFFUSE statement will be inserted into the input file:

```
# FIELD OXIDE

DIFFUSE TIME=100 TEMP=850 T.FINAL=1060
WETO2 PRESS=1.00 \ HCL.PC=0
C.ARSENIC=9.0E19 C.PHOSPHOR=4.0E20
```

Several other parameters not included on the menu are available in the DIFFUSE statement (Chapter 6: “Statements”, Section 6.15:“DIFFUSE”). The DUMP, DUMP.PREFIX, and NO.DIFF parameters can be useful. DUMP and DUMP.PREFIX can be used to make a movie using TONYPLOT. The NO.DIFF parameter specifies that impurity redistribution will be neglected. This provides a good approximation for low temperature processes, such as silicidation.

Several other model specification statements are important for diffusion processes. These are as follows:

- IMPURITY, INTERSTITIAL, and other impurity and point defect statements, which specify model parameters (e.g., diffusivity or segregation) of these species.
- The OXIDE statement, which specifies parameters for different oxidation models.
- The MATERIAL statement, which specifies some basic parameters for all materials.
- The SILICIDE statement, which specifies silicidation coefficients.

Table 7.7 shows the basic diffusion and oxidation models.

Table 7.7. Basic Diffusion and Oxidation Models.

Process	Model	Assumption	Recommendation
Diffuse	Fermi-Default	Defect in equilibrium	For undamaged substrates in inert ambients

	two.dim	Transient defect diffusion	during oxidation, and after medium dose implant (e.g., OED)
	full.cpl	Defect and impurity binding energy model	Post high dose implant & co-diffusion effects, but execution time is high
Oxidation	Vertical	Planar	1D oxidation only (should never be used)
	Compress-Default	Non-planar with linear flow	2D oxidation (e.g. birds beak)
	ViscousElastic	Non-planar with non-linear flow	2D oxidation (e.g. birds beak with thick Si3N4, however, execution time is higher

For a detailed description of all diffusion and oxidation models, see Chapter 3: “SSUPREM4

Models”, Sections 3.1: “Diffusion Models” and 3.3: “Oxidation Models”.

SSPD_Chapter 6_Part 7_Introduction to ATHENA 7_Process Simulator
SSPD_Chapter 6_Part 7_Introduction to ATHENA 7 is the concluding part of this topic 'Introduction to ATHENA'. This module concludes this section with Epitaxy Simulation.

SSPD_Chapter 6_Part 7_Introduction to ATHENA 7_Process Simulator

7.7 Background of ATHENA.

7.7.9 Simulating the Epitaxy Process.

ATHENA/SSUPREM4 can simulate a high temperature silicon epitaxial processes. The epitaxy process is considered as a combination of deposit and diffuse processes. Therefore, processes such as autodoping from a highly doped buried layer into a lightly doped epitaxial layer can be simulated. Diffusion parameters for epitaxial silicon, however, are considered the same as for single crystal silicon.

The epitaxy process is defined in the ATHENA Epitaxy Menu (Figure 7-39). To open this menu, select Process→Epitaxy in the Commands menu. The ATHENA Epitaxy Menu consists of five sections:

- The Time/temperature section selects temperature step parameters in the same way as in the DIFFUSE statement.

- The Thickness/rate section selects either the total thickness of the epitaxial layer, or the deposit rate in microns/minute. In the latter case, the total thickness will be determined by the rate and time.

- The Grid Specification section specifies the vertical grid structure within the grown epitaxial layer. All grid parameters are equivalent to those of the ATHENA Deposit Menu

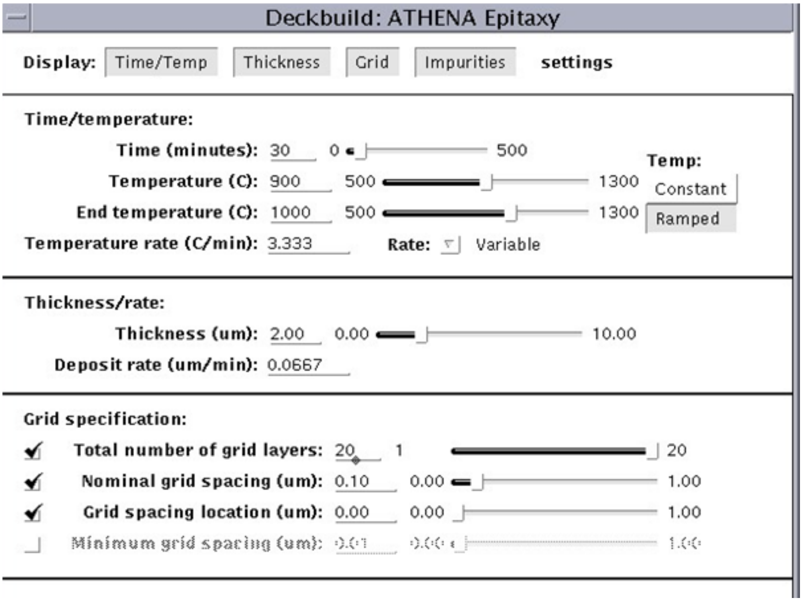
- The Ambient section is where the gas pressure can be modified to the value used in the Epitaxial Chamber.

- The Impurity Concentrations section specifies the growing epitaxial layer in the same way as in the DIFFUSE statement.

All parameters in the last three groups are optional. If the parameters of an epitaxial step are set exactly as shown in Figure 7.39, the following statement will appear in the input file:

```
# EPI-LAYER

EPITAXY TIME=30 TEMP=900 T.FINAL=1000
THICKNESS=5 DIVISIONS=20 \ DY=0.05
YDY=0.00
```

Note: The diffusion during the epitaxy process will use the Diffusion Model Set in the most recent *METHOD* statement. If you need another *METHOD* statement, include it before the *EPITAXY* statement.



Time/temperature:

Time (minutes): 30   0 – 500

Temperature (C): 900   500 – 1300

End temperature (C): 1000   500 – 1300

Temperature rate (C/min): 3.333   Rate: Variable

Temp: Constant / Ramped

Thickness/rate:

Thickness (um): 2.00   0.00 – 10.00

Deposit rate (um/min): 0.0667

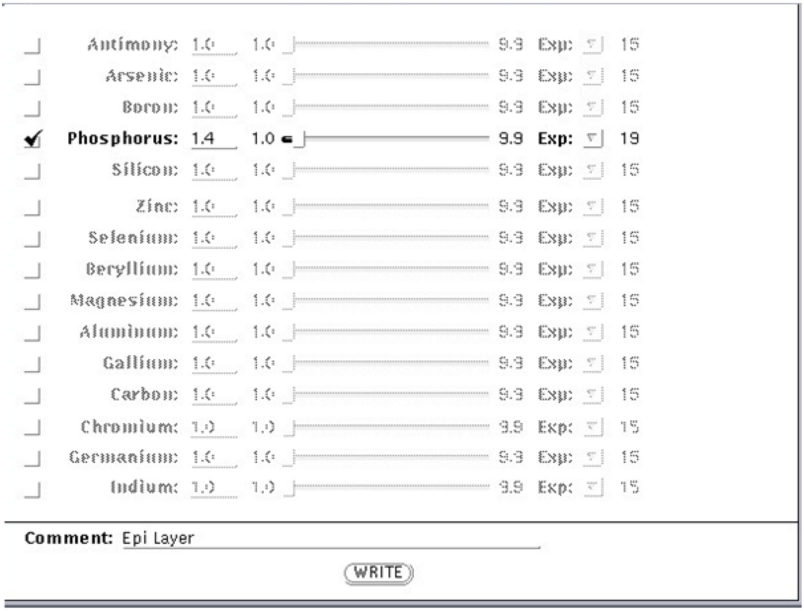
Grid specification:

☑ Total number of grid layers: 20   1 – 20

☑ Nominal grid spacing (um): 0.10   0.00 – 1.00

☑ Grid spacing location (um): 0.00   0.00 – 1.00

☐ Minimum grid spacing (um): 0.01   0.00 – 1.00



☐ Antimony: 1.0   1.0 – 9.9   Exp: 15

☐ Arsenic: 1.0   1.0 – 9.9   Exp: 15

☐ Boron: 1.0   1.0 – 9.9   Exp: 15

☑ Phosphorus: 1.4   1.0 – 9.9   Exp: 19

☐ Silicon: 1.0   1.0 – 9.9   Exp: 15

☐ Zinc: 1.0   1.0 – 9.9   Exp: 15

☐ Selenium: 1.0   1.0 – 9.9   Exp: 15

☐ Beryllium: 1.0   1.0 – 9.9   Exp: 15

☐ Magnesium: 1.0   1.0 – 9.9   Exp: 15

☐ Aluminum: 1.0   1.0 – 9.9   Exp: 15

☐ Gallium: 1.0   1.0 – 9.9   Exp: 15

☐ Carbon: 1.0   1.0 – 9.9   Exp: 15

☐ Chromium: 1.0   1.0 – 9.9   Exp: 15

☐ Germanium: 1.0   1.0 – 9.9   Exp: 15

☐ Indium: 1.0   1.0 – 9.9   Exp: 15

Comment: Epi Layer

Figure 7.39. ATHENA Epitaxy Menu

SSPD_Chapter 6_Part 8_Caliberating ATHENA for typical MOSFET FLOW

SSPD_Chapter 6_part 8 deals with setting the process parameters for simulating MOSFET device.

SSPD_Chapter 6_Part 8_Caliberating ATHENA for typical MOSFET FLOW

7.8 Caliberating ATHENA for typical MOSFET FLOW.

This section of the manual provides information on which parameters should be tuned in the input file to provide predictive simulations using a typical MOSFET process flow. We assume you are now familiar with the mechanics of making an input file and using the correct methods and models (see Section 7.7.3:“Choosing Models In SSUPREM4”). For example, incorrect use of the METHOD statement will invalidate the rest of the following section.

Calibrating an ATHENA input file for a typical MOS process flow involves using the device simulator, ATLAS, since electrical measurements from the MOSFETs in question often represents the majority of the physical data available for calibration. This can be thought as a paradox since ATLAS would also have to be correctly calibrated. The reason that this doesn’t present a problem is discussed below.

An important point to remember when using

Technology Computer Aided Design (TCAD) is that the most critical task is to accurately model the process flow.

Note: For accurate MOSFET simulation, you should invest 90%of the time in achieving an accurate process simulation, while only investing 10%of the time in fine-tuning the device simulation.

The reason for this, especially for silicon technologies, is that the device physics, in general, is understood. For silicon, not only is the physics well understood, it is also well characterized, so most of the default values in ATLAS will be correct. Therefore, the calibration of an ATHENA process file does not involve the calibration of well known quantities such as diffusion coefficients. Instead, the calibration involves variables that are process and production line dependent. For example, the damage caused by an implant cannot be determined exactly, since it is dose rate dependent and can be influenced by beam heating of the substrate, which is dependent on the carousel rotation speed and the efficiency of the cooling system.

Note: If the process has been correctly modeled, the device simulation will also be accurate if appropriate models have been chosen.

If a simulated device exhibits electrical characteristics that are totally inaccurate, you may have done something wrong in the process simulation. Do not make the mistake of changing well known default values in the simulators to make a curve fit one set of results because this will lead to poor predictive behavior. Try and find the cause of a discrepancy.

7.8.1: Input Information

It may seem obvious but must be emphasized that an accurate process flow is vital for simulation accuracy, especially for Rapid Thermal Anneals (see Section 7.7.8:“Simulating Rapid Thermal Anneals (RTA) Notes” for details). Other process information required is an accurate cross-section of the oxide spacer. Modeling the spacer profile accurately ensures the lateral damage distribution due to the subsequent source-drain implants is correctly modeled.

Turning to electrical data, the most important device electrical data is a plot of threshold voltage versus gate length for the NMOS devices. Figure 7.40 shows typical plots of threshold voltage versus gate length. In this figure, the RTA anneal temperature and times were varied to show the various profiles that can be expected. A more typical plot is represented by the 1000°C RTA profile, showing a peak value around 1-2 microns with a tail off for longer or shorter gate lengths.

Figure 7.40: A plot of Threshold Voltage vs. Gate Length for NMOS devices(given in next module)

Gate oxide thickness measurements are also required. Be careful here if oxide thickness is measured with capacitance-voltage (C-V) methods, since quantum effects in very thin oxides (less than 5nm) can lead to inaccuracies because the actual location of the peak concentration of the accumulation charge is not at the interface as classic physics predicts but a short distance into the silicon. Use the QUANTUM model in ATLAS to match accumulation capacitance with oxide thickness for very thin oxides.

Other useful electrical input information is data that won’t be used now but later for the calibration process itself, testing the predictive nature of the simulation. Typical device characteristics used for predictive testing includes threshold voltage versus gate length measurements for a non-zero substrate bias.

7.8.2: Tuning Oxidation Parameters

During oxidation, interstitials are injected into the silicon substrate by the advancing interface. The first parameter to tune is the fraction of consumed silicon atoms that are re-injected back into the substrate as interstitials. In ATHENA, the related tuning parameter is called THETA.0 and is defined in the INTERSTITIAL statement. THETA.0 has been found to be slightly different for wet and dry oxides. The default value is reasonably accurate for dry oxides but some tuning may be required for wet oxidation.

The major effect of interstitial injection during gate oxidation is to create enhanced diffusion of the threshold adjust implant. The measured threshold voltage of the final device is very sensitive to the dopant concentration near the silicon-gate oxide interface. Consequently, threshold voltage measurements are a sensitive indicator of interstitial behavior. Oxidation, however, is not the only source of interstitial injection. The source-drain and LDD implants also induce a large concentration of interstitials. In order to isolate oxidation enhanced diffusion, the threshold voltage of a long gate length device is used, preferably where $L = 20$ µm or more, so that the threshold voltage will be little influenced by damage near the source-drain regions.

Interstitials injected by source-drain implant damage can travel up to 10 µm along the surface before recombination takes place. A gate length of 20 µm is recommended as the minimum gate length for calibration so this can allow the interstitials to diffuse 10 µm along the surface from both the source and drain ends without effecting diffusion near the center of the device. In summary, tuning THETA.0 involves the comparison of modeled and measured threshold voltage data for a long gate-length device.

THETA.0 can be rapidly tuned by taking a one dimensional (1D) vertical cutline through the center of the gate and doing a 1D process simulation. You can either tune THETA.0 manually or by using the Optimize function in DECKBUILD. Theta.0 is tuned until the measured and simulated data of the long channel threshold voltage correspond. The fine tuning of THETA.0 is performed by using a full 2D simulation.

Figure 7.41 shows a typical dependence of extracted threshold voltage on the Theta.0 tuning parameter. Realistic values of THETA.0 correspond to the rising part of the curve. The glitch in the curve is due to rounding errors in the EXTRACT statement used to calculate the threshold voltage due to the automatic and independent mesh generated in the EXTRACT statement. The mesh can be changed from its default value shown here to eliminate this effect. But close examination reveals that the error is only a few millivolts off, which is accurate enough for most process parameter extractions.

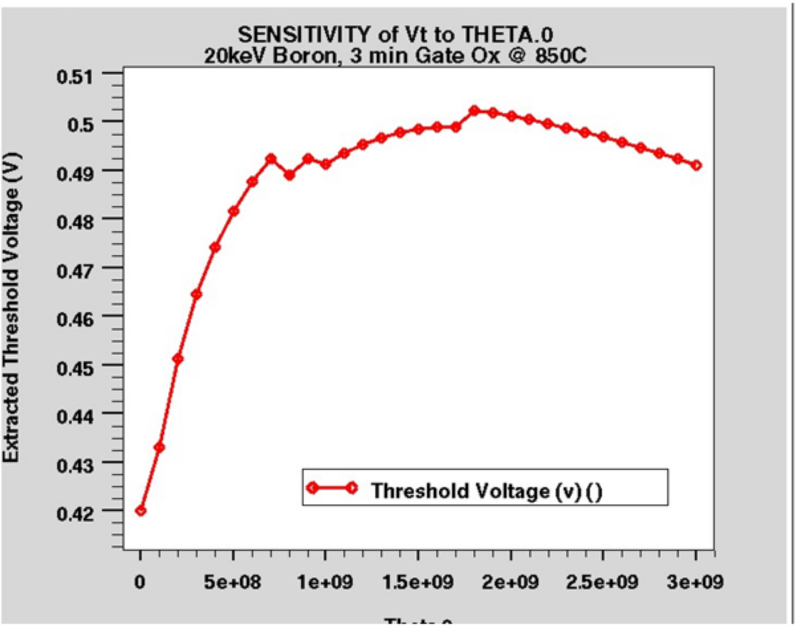


Figure 7.41: A Typical Dependence of Extracted Threshold Voltage on Theta.0

X-axis = Theta.0, Y-axis = Threshold Voltage.

7.8.3: Tuning Implantation Parameters

You can now tune two implantation parameters by using the threshold voltage versus gate length data. The peak value of threshold voltage for a given process flow (the reverse short channel effect) will be a function of the initial implant damage caused by the LDD and source-drain implants. Since these implants have a high total dose and damage, the tuning parameter here is the clustering factor. In ATHENA, this parameter is called CLUST.FACT and is defined in the CLUSTER statement. The higher the clustering factor, the greater the damage, and the greater the diffusion, the greater the reverse short channel effect.

Figure 7.42 shows the effect on the threshold voltage of changing the CLUST.FACT parameter for a typical process flow.

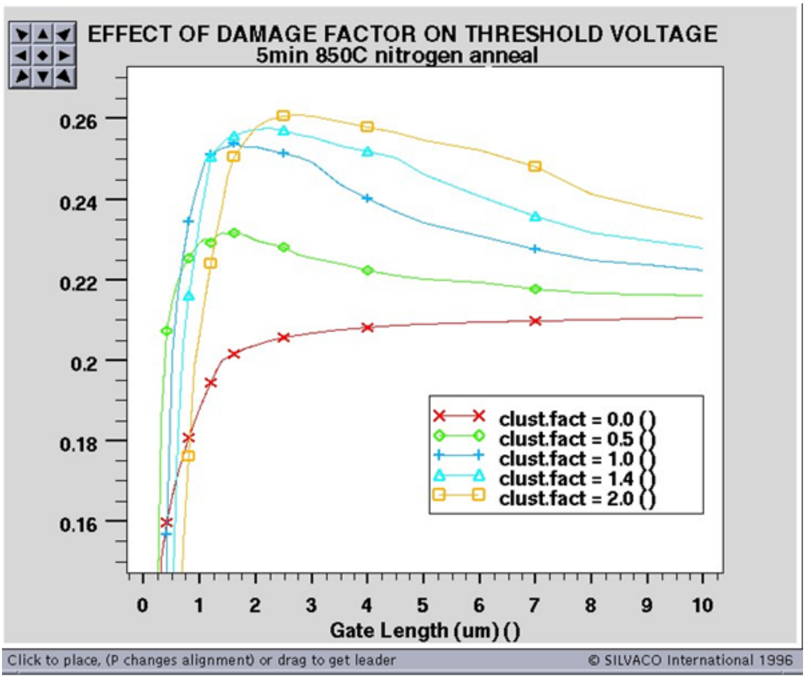


Figure 7.42: How Changing the clust.fact parameter affects the threshold voltage

The second implantation parameter that can now be tuned is the lateral spread of the implant near the surface. In ATHENA, this parameter is called

LAT.RATIO1 and is defined in the IMPLANT statement. The lateral spread of the source-drain and LDD dopant is responsible for the classical short channel effect, where the threshold voltage reduces for very short channel lengths. Simply tune the LAT.RATIO1 parameter until the onset of classical short channel effects of simulated and measured data correspond. If the LAT.RATIO1 is increased, the onset of the classical short channel effect will occur for longer gate lengths.

7.8.4: Tuning Diffusion Parameters

The final part of the threshold voltage versus gate length curve can now be used to tune the surface recombination rate of interstitials. In ATHENA, this parameter is called KSURF.0 and is specified in the INTERSTITIAL statement. The surface recombination of interstitials will dictate the roll-off rate of threshold voltage from its peak value (reverse short channel effect) to the long gate length value. Once again, simply tune KSURF.0 until the long channel threshold voltage roll off rate matches that of the measured data.

PMOS Tuning

PMOS devices are a special case since the boron doped Source/Drain implants overall tend to absorb interstitials rather than emit them. The reverse short channel effect in buried channel PMOS devices can be caused by high angle implants. If high angle implants are used, the reverse short channel effect can be tuned using the LAT.RATIO1 parameter in the IMPLANT statement.

7.8.5: Related Issues on using the Device Simulator ATLAS for MOS Process Tuning

It should now be known that calibrating an ATHENA process file involves using the device simulator ATLAS to a significant extent. Hence, it’s imperative that the use of the device simulator doesn’t create additional errors, rendering the process calibration results invalid.

Fortunately, the device physics involved in simulating the conditions required to extract a threshold voltage are not demanding. The drain voltage required to extract a threshold voltage is only 50-100mV so effects such as impact ionization can be neglected. The field perpendicular to the gate is also relatively low around the threshold voltage so field effects in this direction will do little effect. We recommend, however, using at least the models SRH and CVT during the calculation. Other parameters for silicon are sufficiently well known for silicon to the point that the results from the device simulator are reliable.

The first important point is to ensure that you let the device simulator calculate the work function of the gate electrode from the simulated doping profile rather than assigning a value to it. This means, making sure that the polysilicon gate is not itself defined as an electrode but rather a layer of metal, usually aluminum, is deposited on top of the polysilicon gate. Therefore, this metal layer is the film defined as the electrode. Do not assign a work function to this deposited metal electrode to ensure that it behaves as an ohmic contact rather than a Schottky contact. The effective work function of the poly gate will then be correctly calculated from the doping profile in the polysilicon.

An important area for accuracy in MOSFETs is modeling the inversion region under the gate. As it is, this charge that is responsible for current conduction in the device. The inversion region charge under the gate-only extends approximately 30 Angstroms into the silicon. The inversion region charge density under the gate falls off rapidly with depth into the silicon. It is imperative that there are several mesh points in the Y direction in this inversion region to model the drain current correctly. Accordingly, we recommend that the mesh spacing under the gate be no more than 10 Angstroms (1 nm).

You would think that a 10 Angstrom mesh under the gate would result in a huge number of mesh points. But, there only needs to be approximately three mesh points within the inversion region in the

Y direction. The grid spacing can increase rapidly in spacing away from the oxide-silicon interface.

Figure 7.43 shows the effects of changing the mesh spacing at the interface on the simulated drain current. You can see from this figure that too coarse of a mesh always results in too high of a current simulated.

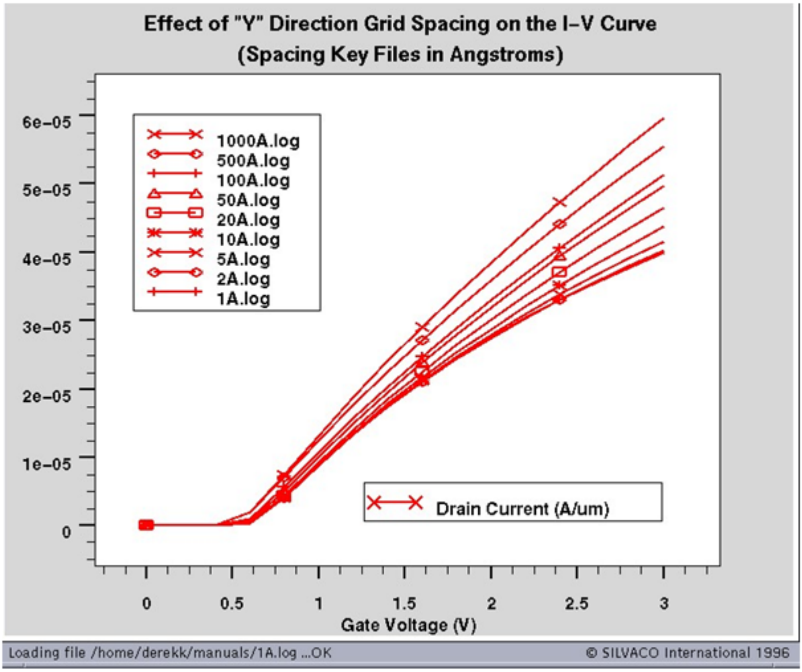


Figure 7.43: The effect of changing the mesh spacing at the interface on the simulated drain current

If contact resistance is a problem, then include it in the CONTACT statement. The resistance added to the CONTACT statement should be the measured resistance per contact divided by the number of contacts on each individual electrode. Obviously for D.C. measurements, the resistance on the gate contact will have no effect on the results since no current flows in this direction.

Checking the Predictive Powers of Tuned Process Parameters

If the process simulation has been correctly tuned, the process and device simulators should have predictive powers. To check the validity of the tuning process, use a new set of electrical data that was not used during the tuning process. For example, a good alternative set of data is to check the threshold voltage versus gate length for a non-zero voltage applied to the MOSFET body contact.

Conclusion

Using just one set of easily obtained measured electrical data, namely a plot of threshold voltage versus gate length, you can obtain most of the tuning parameters required for accurate process simulation. The other most important piece of data required is an accurate measurement of the gate oxide thickness, which is routinely measured in any instance.

You have been given specific advice as to which process and device models to use for each process in order to get the best results out of the simulation software. In particular, the correct use of models for the implantation and diffusion processes is stressed, as this has a dramatic effect on MOSFET characteristics, especially as anneal times and device dimensions decrease.

SSPD_Chapter 6_Part 9_Caliberating ATHENA for typical Bipolar Flow Process

SSPD_Chapter 6_Part 9_Caliberating ATHENA for typical Bipolar Flow Process. For this module a good understanding of the Physics of BJT is must.

SSPD_Chapter 6_Part 9_Caliberating ATHENA for Bipolar Process.

7.9. Calibrating ATHENA for a Typical Bipolar Process Flow.

As with MOS calibration text, we assume you are familiar with the mechanics of making an input file and using the correct methods and models (see Section 2.4:"Choosing Models In SSUPREM4"). For example, incorrect selection of diffusion models defined in the METHOD statement would invalidate the remainder of the following section.

Calibrating a bipolar process flow entails matching the two parameters, base current and collector current versus base emitter voltage to measure results throughout the full operating range of the device. By implication, the current gain of the device (Ic/Ib) will also be matched. All of the following paragraphs refer to the standard plot of collector and base currents measured against the base-emitter voltage, Vbe, unless it's specifically stated otherwise. This standard I-V graph is usually referred to as the Gummel Plot.

Another way of plotting the same information in a different format that can prove useful is a plot of current gain, hfe, versus the log of the collector current. This graph, however, is a derivation of the same information that makes it less clear as to which current is increasing or decreasing for each change. Therefore, a less useful graph when it comes to understanding exactly what is happening to the collector and base currents.

The full operating range of a bipolar junction transistor (BJT) consists of three general regions defined by the current density injected into the base. These three operating regions are usually described as low, medium, and high current injection regimes. The medium injection region is the most important part of the curve to model correctly as this represents the typical operating condition of the BJT. Each of the three operating regions is dominated by a different physical phenomenon. Therefore, successful modeling of a BJT involves matching both the base and collector currents in each of the three general operating regions, making a total of six areas for calibration. The derived parameter, hfe, is also a good parameter to monitor, since this is sensitive to errors in the ratio of collector to base current.

The following text suggests an approach and describes which of the six regions are effected by each change. The general technique is to calibrate the parameters that have the greatest effect on device performance in all regions first and then to move on to more subtle phenomenon that effect certain parts of the base or collector currents or both. In general, matching the collector current for all injection regions is less problematic than matching the base current at the extremes of the injection regions. Consequently, there are more sections on tailoring these parts of the curve. The text is divided into the following sections:

1. Tuning Base and Collector Currents – All Regions

2. Tuning the Base Current – All Regions

3. Tuning the Collector Current – All Regions

4. The Base Current Profile – Medium Injection

5. The Base Current Profile – Low Injection

6. Conclusions

If you follow this order, there should be a reasonable correlation between measured and simulated data. Most of the tuning parameters, however, have some degree of interdependency to the extent of which is also device design specific. Therefore, some degree of iteration of the tuning parameters is to be expected.

When tuning bipolar transistors, there is a greater emphasis to accessing tuning parameters by using the device simulator, ATLAS, compared to optimizing MOSFETs, where most tuning parameters are process-related. A powerful combination is the tuning of a BiCMOS process where you can use the MOSFET part of the process flow to tune the process parameters, while using the Bipolar part of the flow to tune ATLAS. This technique should yield a high degree of predictability in the results.

Tuning the process simulator parameters in ATHENA are mainly required to model effects, such as the implantation induced defect enhanced diffusion responsible for the Emitter Push Effect, which is essential to obtain the correct depth of the base-collector junction. The correct process modeling of the out diffusion of dopant from the poly-emitter into the mono-crystalline substrate is also critical to obtaining well-matched I-V curves. Another critical process modeling area is the base implant, because it is essential to match measured and modeled base resistance for correct modeling of the collector current. These and other issues are discussed in these sections.

7.9.1: Tuning Base and Collector Currents – All Regions

The most important parameter to model the general level of base and collector currents is the device measurement temperature. The base and collector currents are strongly influenced by temperature changes, as small as a few degrees centigrade. A significant effort should be made to determine the exact temperature of the device during measurements before calibration is attempted. This temperature should be input into ATLAS in the MODELS statement using the TEMPERATURE=<n> parameter. An increase in temperature will cause an increase in base and collector currents.

7.9.2: Tuning the Base Current – All Regions

A critical region for poly-emitter bipolar devices is the interface between the poly-emitter and the mono-crystalline silicon. This region is difficult to process simulate directly as the interface between the polysilicon emitter and single crystalline silicon usually consists of a thin, uneven and possibly non-continuous film of oxide. This is simulated by calibrating the overall effect of this interface with ATLAS. The tuning parameter is the surface recombination velocity at this interface for electrons (VSURFN for PNP devices) or holes (VSURFP for NPN devices). This will only be effective for thin emitters where at least a fraction of the holes (for NPN devices) can reach the emitter before recombination.

The surface recombination velocity parameter not only affects the base current, it also affects the base current in all of the operating regions. Therefore, it is a powerful parameter to approximately match the base current and gain throughout the full operating range. In some cases, the base current may be less affected in the very high and very low injection regions by changes in the surface recombination velocity, and adding some scope to fine tuning the profile of the base current versus base-emitter voltage curve.

It is important to define the poly-emitter as an electrode so it can define the interfacial surface recombination velocity, VSURFN and VSURFP, using the CONTACT statement. This is in contrast to the MOSFET calibration text where we strongly advise you not to define the polygate as an electrode. Be sure not to get these two confused. The parameter that activates the recombination velocity is SURF.REC, which is also in the CONTACT statement. For example, an NPN BJT statement would be:

```
CONTACT NAME=emitter N.POLYSILICON
SURF.REC VSURFP=1.5e5
```

A lower value of recombination velocity, VSURFP, will reduce the base current and increase the gain, hfe. The reverse is also true.

7.9.3: Tuning the Collector Current – All Regions

Figure 7.43 shows the parameter that affects the collector current over the entire range is the intrinsic base resistance. The base resistance is primarily determined by the dose of the base implant(s). An increase in the base implant dose will decrease the intrinsic base resistance and decrease the collector current in all injection regions. In some cases, however, the collector current may be affected a little in the very high injection region, giving scope for fine tuning the profile of collector current versus base- emitter voltage.

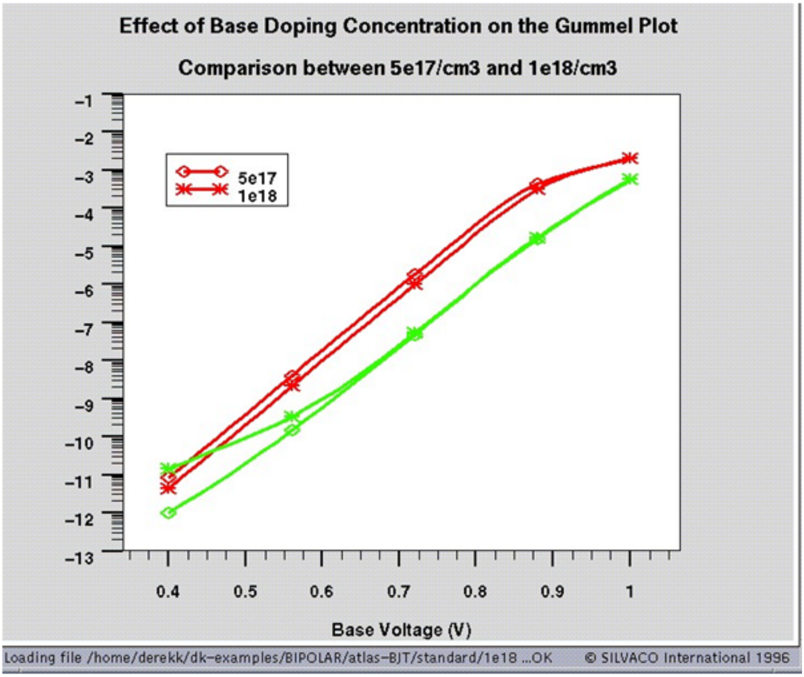


Figure 7.43: Effect of base doping profile on low injection base current in BJT

If the pinched or intrinsic base sheet resistance is a measured parameter, the simplest way to match measured and simulated data is to make slight changes to the base implant dose so that the simulated dose is not outside the expected error in actual implanted dose in conjunction with the error in percentage activation.

In some designs, where the base contact is close to the collector contact or the base contact is the substrate or is generally wide, the collector current can also influence all current injection regions by specifying a surface recombination velocity at the base contact. For a typical design with a buried n+ collector and surface contacts, the surface recombination velocity at the base contact may have little affect on the collector current.

7.9.4: The Base Current Profile – Medium Injection

In ATLAS, there are two major parameters that have a significant affect on the base current in the medium injection regime. These parameters are the Poly-emitter Work Function and the Bandgap Narrowing Effect. These parameters are described below.

Poly-emitter work function

If the poly-emitter is described as N.POLYSILICON in the CONTACT statement for an NPN device, as already described, the Poly-emitter Work Function is then set to 4.17 V and is correct for saturation doped n++ polysilicon. But if the poly-emitter is not saturation-doped, the work function will differ from this ideal and have a pronounced affect on the base current and current gain in the medium injection regime as shown in Figure 7.44. The work function of the poly-gate can vary from 4.17 V for n++ poly-silicon to (4.17 V + Eg) for p++ polysilicon, depending on the position of the Fermi-Energy. Changing the work function of the poly-emitter by just 0.1 V from 4.17 V to 4.27 V can often reduce the current gain in half in the medium injection regime, so it’s very important to assign the correct value. The CONTACT statement below assigns a work function of 4.27 eV to the poly-emitter, while keeping the other parameters the same as before.

```
CONTACT NAME=emitter SURF.REC
VSURFP=1.5e5 WORKFUN=4.27
```

The poly-emitter work function can be calculated by measuring the position of the Fermi-Energy at the poly-silicon/silicon interface relative to the conduction band and adding this value to 4.17V. For example, if the Fermi-Energy is measured as being 0.1eV from the conduction band edge, the work function of the poly-emitter set in the CONTACT statement should be set to 4.17 + 0.1 = 4.27V.

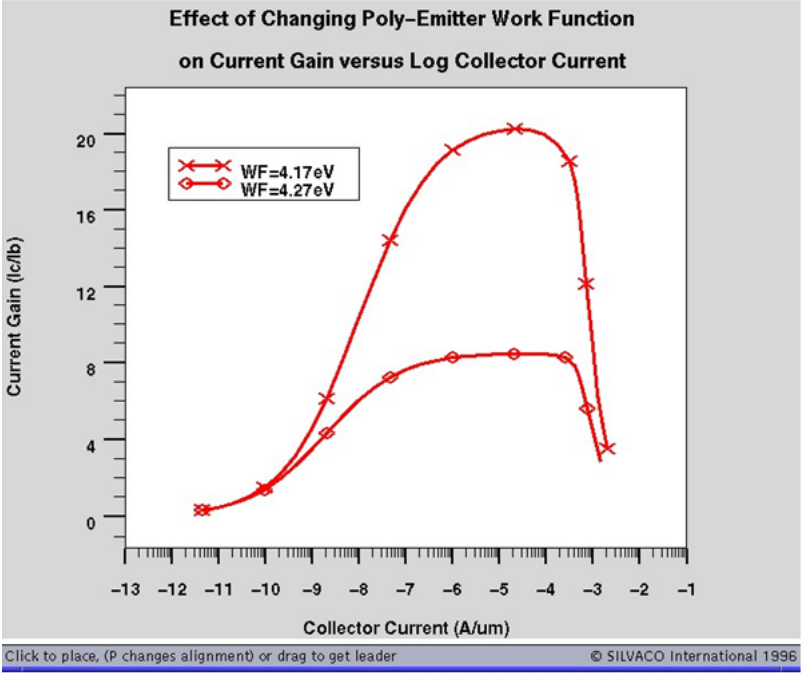


Figure 7.44: Effect of emitter contact work function on bipolar gain

Bandgap Narrowing Effects

If the BIPOLAR parameter is stipulated in the MODELS statement in ATLAS, bandgap narrowing is included automatically. The inclusion of bandgap narrowing in the MODELS statement is strongly advised since this phenomenon has a significant effect on the current gain of the device. But, to validate the default Klaassen bandgap narrowing model, you should also use the Klaassen mobility model. Use the additional keyword KLA in the MODELS statement to activate this model. For example:

```
MODELS BIPOLAR KLA
```

The parameters in the Klaassen bandgap narrowing model are user-definable in the MATERIAL statement and described in the “Physics” Chapter of the ATLAS USER’S MANUAL, VOL. I. There are three user-definable parameters for the Klaassen band gap narrowing model. The BGN.E parameter has a linear dependency on doping concentration and has the default value of 6.92e-3 volts. BGN.C has a square root dependency with doping concentration and has the default value of 0.5. BGN.N is the value of doping where band gap narrowing effectively starts to take effect and has a default value of 1.3e17/ cm3. The equivalent default setting consequently should be written as:

```
MATERIAL BGN.E=6.92e-3 BGN.C=0.5
BGN.N=1.3e17
```

You can alter these parameters to modify the current gain of the device in the medium injection regime. For example, reducing the linear parameter from 6.92e-3 to 6.5e-3 is sufficient to cause a significant increase in current gain in the medium injection region. Although the bandgap narrowing parameters affect both collector and base currents, the base current is affected to a greater degree. The most sensitive plot to see the effect of small changes to bandgap narrowing is a plot of current gain versus log of collector current. A reduction in bandgap narrowing will result in an increase in current gain in the medium current injection region.

7.9.5: The Base Current Profile – Low Injection

This is one case where there is an interdependency on one parameter, since the intrinsic base resistance not only affects the collector current in all regions (see the previous section) Figure 7.43, however, also has an effect on the base current in the low injection region.

For a small range of implant doses around the optimum, the base doping concentration will also affect the position of the knee or the rate or both of fall off of the base current in the low injection operating region of the device. This is most noticeable as a loss of current gain in the low injection region for the alternative standard plot of current gain versus collector current. An increase in the base implant reduces the intrinsic resistance and typically increases the base current in the low injection region, resulting in a decrease in current gain for very low currents.

A similar effect to increasing the base doping is observed if the base doping is kept constant but the overall doping is reduced in the mono-crystalline silicon region of the emitter. You can tune the doping profile in the mono-crystalline region of the emitter using three parameters in ATHENA. The main physical effect of these ATHENA parameters is to change the doping profile of the emitter in the mono- crystalline silicon. These process parameters are as follows:

- The total interstitial concentration in the poly-emitter.

- The dopant segregation effects in the poly-emitter.

- The dopant velocity across the silicon/polysilicon boundary.

The first process parameter will affect how quickly the dopant in an implanted poly-emitter reaches the silicon/polysilicon boundary during the RTA diffusion and therefore affects the total diffusion of dopant into the single crystalline part of the emitter and the base width doping profile.

The second process parameter affects dopant pile-up at the poly-silicon/silicon boundary and therefore the source doping concentration at the mono-crystalline interface. Once again, this will affect the overall doping profile of the emitter in the mono-crystalline region of the device.

The third process parameter affects the velocity of transport of dopant across the polysilicon/silicon boundary with similar effects to the parameters above.

You can use these parameters to tailor the emitter doping profile in the mono-crystalline silicon region to match available measured data, usually in the form of SIMS or capacitance information. An accurate profile of dopant in the poly-silicon part of the emitter is not too important if measured data concerning interfacial dopant concentrations is available. This is because the work function of the poly-emitter will be set in ATLAS by defining the poly-emitter as an electrode. All you need to calculate the correct work function at the poly-silicon emitter is the interfacial doping concentration at the poly- silicon/silicon interface on the poly side of the junction. See the “Poly-emitter work function” Section on page 2-51 for setting the correct work function for the poly-emitter .

Conclusions

By using a logical combination of tuning parameters available in both the process simulator (ATHENA) and the device simulator (ATLAS) and with the influence of each parameter, you can get a good match for bipolar transistors for most device designs.

Since it is usually less problematic to match the collector current for all levels of applied base-emitter voltage compared to the matching of base current, you will probably find that more time is spent trying to match the base current for very small and very large values of applied base-emitter voltage. You should, however, spend a good amount of time on making sure that the correct process models are used in the process flow to reduce the overall uncertainty as to which parameters require calibration.

SSPD_Chapter 6_Part 9_Appendix Device Physics

SSPD_Chapter 6_Part 9_Appendix on Device Physics describes the models used in ATLAS and ATHENA

SSPD_Chapter 6_Part 9_Appendix Device Physics

Section I. The Bipolar Junction Transistor Technology in 60's-70's.

In 1959 at the time IC Technology was invented, the vertical NPN transistor looked as given in Figure I.

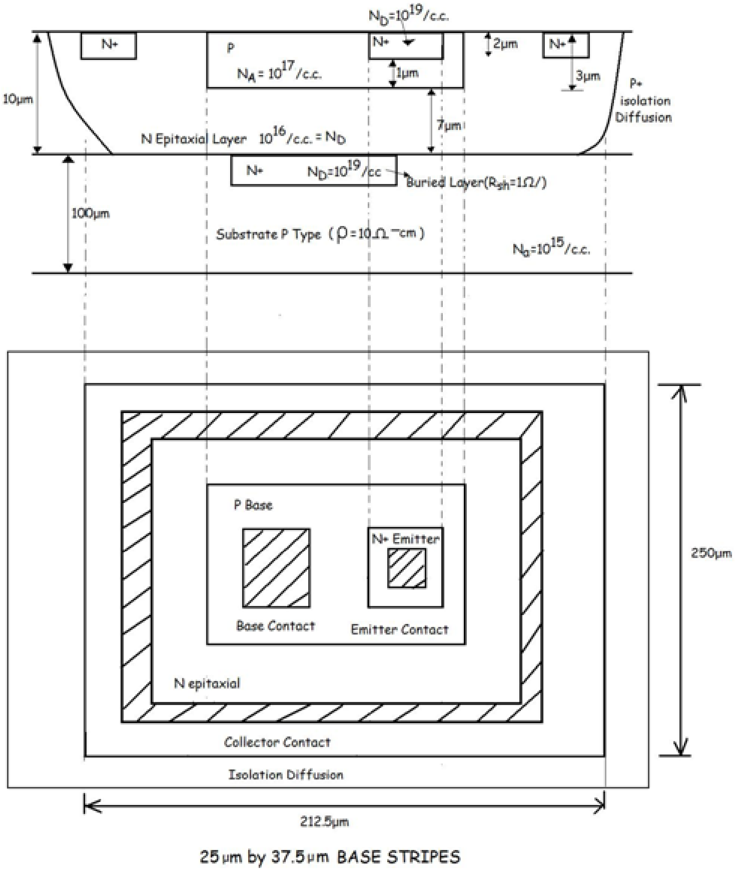


Figure I. Cross-sectional view and Plan view of Vertical NPN transistor.

(25um by 37.5um base stripe geometry)

Total area covered 212 μm **by 250** μm

Substrate is 100um thick.

Epitaxial layer is 10um thick.

Base Collector Junction is at 3um depth and Emitter-Base Jn. at 2um depth.

Hence Base Width = W_B = 1 um thick.

P Substrate- $\rho = 10\ \Omega$-cm, $N_A = 10^{15}$ Boron atoms/cc;

Buried Layer has been put to reduce Collector Series Resistance.

$N_D = 10^{19}$Arsenic Atoms/cc; In buried layer Arsenic is used as donor type dopent because it has much lower diffusion coefficient hence almost negligible out diffusion in subsequent diffusion and heating cycles.

N Epitaxial layer- $\rho = 0.1\ \Omega$-cm, $N_D = 10^{16}$ Phosphorous Atoms/cc;

Base layer- $N_A = 10^{17}$ Boron Atoms/cc; Base layer has a sheet resistance of 200 Ω/sq.

Emitter and Collector Contact layer- $N_D = 10^{19}$ Phosphorous Atoms/cc, $R_{sh} = 1\Omega$/sq;

Section II. The Bipolar Junction Transistor Technology in 1980’s-2000’s.

Gordon Moore is a co-founder of Intel and he made an empirical observation which became a Law . It stated :

“ at our rate of technological development, the complexity of an integrated circuit, with respect to minimum component cost will double in about 24 months”.[“Cramming more components onto Integrated Circuits”, *Electronics Magazine,*19 April 1965]. The law has held the test of time to date as is evident from Table I.4_50 years journey of IC Technology_Appenidix I.

Today Moore’s Law has become a self fulfilling prophecy and a goal Industry tries to achieve all the time. In attempting to maintain this rate of growth, enormous R&D has gone into the development of tools and equipments required for IC fabrication and into the advancement of processing techniques required for IC fabrication.

Section III.1. The relation between the physical parameters and performance parameters and Physics of High Performance nano BJT.

In CB BJT:

$$I_C = \alpha_F M I_E + M I_{CBO} \qquad I$$

Here $\alpha_F = \gamma \times \beta^*$

= (Emitter Injection Efficiency) × (Base Transport Factor)

Current Transfer Ratio

$$= \frac{I_{En}}{I_E} \times \frac{I_{Cn}}{I_{En}} \times \frac{I_C}{I_{Cn}} = \gamma \times \beta^* \times M \quad II$$

In Eq. I , M = Avalanche Multiplication Factor =

$$\frac{1}{1-(\frac{V_{CB}}{BV_{CBO}})^n}$$

III

Where n is Miller Indices ranging from 2 to 7.

At normal Collector to Base Voltages, M = 1

But as V_{CB} becomes large approaching BV_{CBO} , M becomes a very large number and eventually it becomes infinite at $V_{CB} = BV_{CBO}$. This is known as CB Junction Breakdown by Avalanche mechanism.

At low doping we have to invoke Impact Ionization Model or Avalanche Mechanism. At doping densities larger than 10^{18} dopent /cc we must invoke mechanism of tunnelling also known as Zener Breakdown Model.

As seen in Figure II, area enclosed by Forward Active Mode bracket is the Forward Active Mode region of operation of CB BJT. In this region the active device behaves as a linear device and is suitable for linear applications. The safe operating voltage range extends from 0 V to (BV_{CBO} – 2 V) in CB BJT.

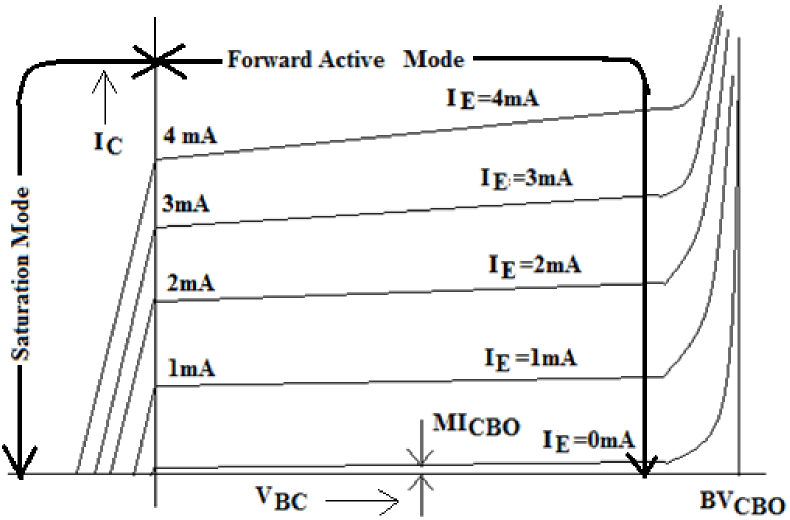


Figure II. CB BJT output family of curves: I_C vs V_{BC}.

In CE BJT, Equation I is rearranged as shown below:

$$I_C = \frac{\alpha_F M I_B}{(1 - \alpha_F M)} + \frac{I_{CBO}}{(1 - \alpha_F M)} \qquad IV$$

At low voltages, Equation II reduces to :

$$I_C = \frac{\alpha_F I_B}{(1 - \alpha_F)} + \frac{I_{CBO}}{(1 - \alpha_F)} \qquad V$$

This further simplifies to:

$$I_C = \frac{\beta_F I_B}{1} + \frac{(1+\beta_F)I_{CBO}}{1} \qquad IV$$

Or

$$I_C = \beta_F I_B + I_{CEO}$$

V

Here I_{CBO} = Collector Current in CB BJT with Emitter Open that is Emitter current is held at zero magnitude. This current is the reverse leakage current across the reverse biased CB Junction and is of the order of nA.

Here I_{CEO} = Collector Current in CE BJT with Base Open that is Base current is held constant. This current is not the reverse leakage current across the reverse biased CB Junction. It is the Collector Current with BJT in Forward Active Mode and it is of the order of μA.

Equation IV indicates that at a given 'V_{CE}*' $\alpha_F M$ becomes UNITY and both the terms in the Equation become infinity. Physically this means that as V_{CE} increases, by Base Width Modulation also known as Early Effect, forward current transfer ratio improves. Simultaneously M is also gradually increasing as seen in Equation III. Hence Collector to Emitter Breakover takes place across Collector to Emitter terminals and this voltage is designated as

BVCEO .

At BVCEO ,

$$\alpha_F M = \frac{\alpha_F}{1-(\frac{V_{CB}^*}{BV_{CBO}})^n} = 1 \qquad VI$$

Rearranging the terms we get:

$$V_{CB}^* = BV_{CEO} = BV_{CBO}(1-\alpha_F)^{1/n} = \frac{BV_{CBO}}{\sqrt[n]{(1+\beta_F)}} \qquad VII$$

Therefore if Collector-Base Breakdown occurs at 128 V then Collector-Emitter Breakover with constant base current drive occurs at 60V as shown in Figure III. The range of voltage over which BJT can be safely operated is drastically reduced. The same CE BJT if used with constant voltage drive, as it is in Current Mirror configuration, will experience breakover at 112V.

The physics of this Breakover Phenomena becomes clear from Figure IV and Figure V.

Figure IV shows the electron and hole component of IE , IC and IB at low voltages when reverse leakage current at CB Junction is:

$$\frac{I_{CBO}}{(1-\alpha_F)}$$

Figure V shows the electron and hole component of I_E , I_C and I_B at high voltages when reverse leakage current at CB Junction is:

$$\frac{I_{CBO}}{(1-\alpha_F M)}$$

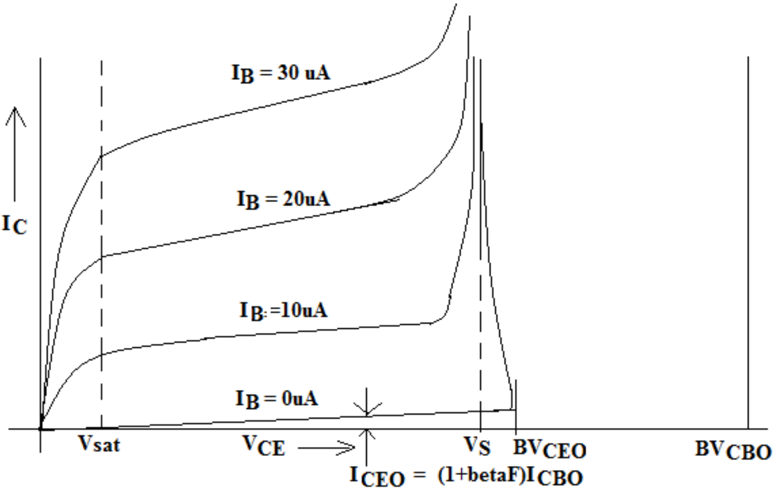


Figure III. CE BJT Output Family of Curves: I_C vs V_{CE} .

Figure IV and Figure VI must be studied in correspondence. Figure VI shows the carrier concentration profile of NPN transistor in Forward Active Mode.

In Forward Active Mode, EB junction is forward biased and BC junction is reverse biased. Built-in barrier potential at EB is reduced $(\varphi_{BO} - V_{EB})$ and built-in barrier potential at BC junction is increased $(\varphi_{BO} + V_{CB})$.

Reduced barrier potential helps in electron(majority carrier) injection into base. In base electron becomes minority carriers hence increased barrier potential at CB junction is down-hill for the minority carriers. Therefore electrons injected into Base transit through the narrow Base region with negligible recombination and are collected by Collector. Since base is narrow practically all the carriers in Base are collected by the Collector.

While electrons are travelling longitudinally across the Vertical NPN BJT they experience recombination in Emitter as well as in Base.

Recombination in Emitter causes Emitter Injection Efficiency (γ) to be less than 100%.

Recombination in Base causes Base Transport Factor (β^*) to be less than 100%.

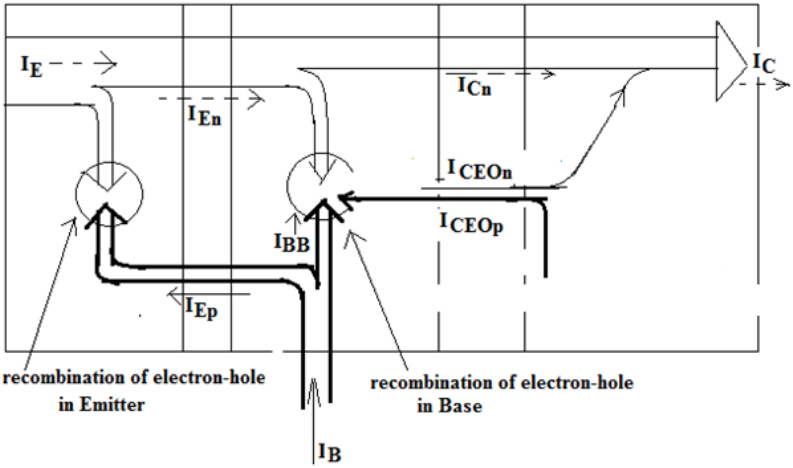

I_E
I_{En}
I_{Cn}
I_C
I_{CEOn}
I_{CEOp}
I_{BB}
I_{Ep}
recombination of electron-hole in Emitter
recombination of electron-hole in Base
I_B

Figure IV.Electron-hole component of I_E , I_C and I_B in CE BJT.
Thick line shows the movement of holes. Thin line shows the movement of electrons.

These two recombinations cause the forward current transfer ratio to be less than UNITY. To provide these two recombinations Base Current is necessitated.

$I_B = I_{BB} + I_{Ep}$ = provides the holes for recombination in Base + provides the holes to be injected into Emitter and cause hole diffusion current in Emitter.

Therefore

$$I_{BB} = \frac{Q_B}{\tau_B} = \frac{electron\ charge\ stored\ in\ Base}{lifetime\ of\ electrons\ in\ Base}$$

VIII

In Figure VI, charge stored in the Base is shown by the shaded region and is given by the following expression:

$$Q_B = \frac{qAW n_p(0)|_B}{2}$$

where

$$n_p(0)|_B = n_{po} Exp\left(\frac{V_{EB}}{V_T}\right) = n_{po}\lambda$$

IX

$$I_{Ep} = qAD_p \frac{p_n(0)|_E}{L_p}$$

where

$$p_n(0)|_E = p_{no}\lambda$$

X

In Figure VI, minority carrier densities are depressed near the depletion layer of BC junction. Here EHP are thermally generated since minority carrier density is less than the thermal equilibrium value. The EHP (electron-hole pair) which survive to reach the reverse biased BC junction depletion width are swept across the depletion width as the reverse leakage current. The electrons on Base side and holes on Collector side are swept across the depletion width because the enhanced built-in barrier potential across the BC junction acts as the down-hill for minority carriers on two sides. The electron component of this reverse leakage I_{CEO} adds to the injected electron current being collected by the collector and hole component of the reverse leakage I_{CEO} provide the holes in Base for recombination as shown in Figure IV.

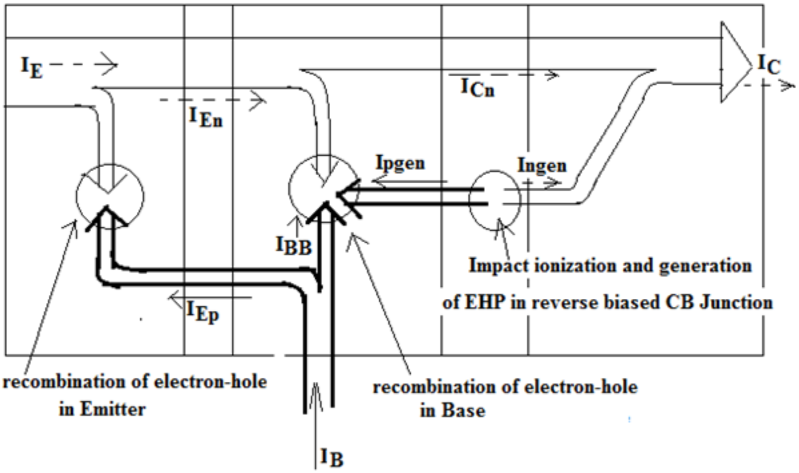

I_E
I_{En}
I_{Cn}
I_C
Ipgen
Ingen
I_{BB}
Impact ionization and generation of EHP in reverse biased CB Junction
I_{Ep}
recombination of electron-hole in Emitter
recombination of electron-hole in Base
I_B

Figure V. Electron -hole component of I_E, I_C and I_B at high voltages when impact ionization and generation starts in CB junction.
Thick line shows the movement of holes. Thin line shows the movement of electrons.

As V_{CE} increases , down-hill potential gradient becomes very steep which accelerates the minority carriers being swept across the BC junction to the extent that IMPACT IONIZATION and IMPACT GENERATION starts as shown in Figure V. Now a copious amount EHP are generated in reverse biased BC junction. Electrons add to I_C increasing it considerably and holes povide recombination centers in Base.

Eventually when $\alpha_F M$ becomes UNITY, holes generated by impact ionization provides all the holes required in Base for recombination with injected electrons as well as holes generated by impact ionization provide the holes to be injected into Emitter. Thus requirement for I_B from Base terminal becomes zero, current gain becomes infinite and I_C shoots up and breakover occurs. At the breakover point S-type negative impedance region is caused as shown in Figure VI. The PHYSICS of BREAKOVER is as follows:

At $\alpha_F M = 1$ the impact generated holes just suffice the recombination needs of the device. At $\alpha_F M > 1$ the impact generated holes are more than what is required for recombination in the Base and for injection into Emitter. Hence the excess part of impact generated holes starts building the concentration of the carriers and their concentration gradient in the Base. This leads to a sudden shoot up in the Collector Current.

The S-type NIR occurs as shown in Figure III, because before breakover α_{F1} is of the order of 0.7 4 because of very small I_C and after the breakover α_{F2} is of the order of 0.8 because of large I_C hence

$$\alpha_{F1}M_1 = \alpha_{F2}M_2$$

$=1$

$$\frac{\alpha_{F1}}{1-\frac{BV_{CEO}}{BV_{CBO}}^{n}} = \frac{\alpha_{F2}}{1-\frac{V_S}{BV_{CBO}}^{n}} \qquad XI$$

Rearranging the terms we get:

$$V_S = BV_{CBO}\sqrt[n]{1-\frac{\alpha_{F2}}{\alpha_{F1}}[1-(\frac{BV_{CEO}}{BV_{CBO}})^{n}]} \qquad XII$$

For 2N2119A: $n = 2$, $BV_{CBO} = 128V$, $BV_{CEO} = 66 V$, $\alpha_{F1} = 0.74$; $\alpha_{F2} = 0.8$.

Substituting these values we get: $V_S = 58V$.

Section III.2. Band Gap Narrowing in BJT.

In a classical BJT, Emitter Bulk Width(2μm) > > Diffusion Length of minority carriers in Emitter(fraction of a micron). Hence hole (minority carrier in Emitter) diffusion current is non-dominant and d.c. current gain in a classical BJT is :

$$\beta_F = \frac{I_C}{I_B} = \frac{I_{En}(electron\ diffusion\ in\ Base)}{I_{Ep}(hole\ diffusion\ in\ Emitter)+I_{base\_recombination}}$$

XIII.

$$I_B = -qAD_p \frac{p_n(0)|_E}{L_p} + \frac{Q_B}{\tau_B} \tag{XIV}$$

$$I_C = I_{En} = -qAD_n \frac{n_p(0)|_B}{W_B} \tag{XV}$$

Therefore in a classical BJT the dc current gain is as given below:

$$\frac{I_C}{I_B} = \frac{-qAD_n \dfrac{n_p(0)|_B}{W_B}}{-qAD_p \dfrac{p_n(0)|_E}{L_p} + \dfrac{Q_B}{\tau_B}} \tag{XVI}$$

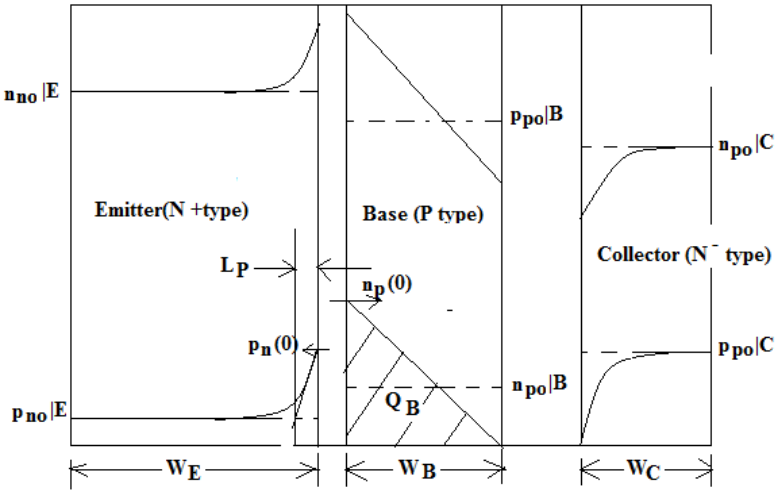


Figure VI Carrier Concentration Profile along the longitudinal axis of the Vertical IC NPN Transistor.

But as lateral scaling and vertical scaling have scaled down in the process of evolution from SSI to MSI to LSI, base width have been scaled down from 1000nm to 100nm and now they are being scaled down to 10nm. Emitter Bulk Width (W_E) have scaled down from 2 microns to a fraction of micron. Hence diffusion length of holes [$L_P = \sqrt{(D_P \tau_P)}$] in Emitter has become comparable to Emitter Bulk Width and carrier concentration profile is akin to that of a narrow base diode. This means two things:

1. In the Expression II, Base Recombination Current becomes negligible compared to hole diffusion current in Emitter;
2. Hole Diffusion current becomes a function of surface recombination velocity at the emitter contact.

At a metal Emitter Contact surface recombination velocity is very high. Hence dc current gain drastically deteriorates.

In Equation XVI , diffusion length is replaced by emitter bulk width and base recombination current is neglected.

$$\frac{I_C}{I_B} = \frac{-qAD_n \frac{n_p(0)|_B}{W_B}}{-qAD_p \frac{p_n(0)|_E}{W_E}}$$

XVII

$$n_p(0)|_B = \overline{n_p}\, Exp\left(\frac{V_{EB}}{V_T}\right) = \frac{n_i^2}{N_A}\lambda = \frac{N_C N_V Exp(\frac{-E_{gB}}{kT})}{N_A}\lambda$$

$$p_n(0)|_E = \overline{p_n}\, Exp\left(\frac{V_{EB}}{V_T}\right) = \frac{n_i^2}{N_D}\lambda = \frac{N_C N_V Exp(\frac{-E_{gE}}{kT})}{N_D}\lambda$$

Where

$$\lambda = Exp\left(\frac{V_{EB}}{V_T}\right)$$

Therefore

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{N_{D|E}}{N_{A|B}}\right) Exp\left[\frac{(E_{gE} - E_{gB})}{kT}\right] \qquad XVIII$$

If we take the following data:

$(N_D)_E = 10^{20}$/cc, $(N_A)_B = 10^{17}$/cc, $W_E = W_B =$ 1µm and $D_{pE} = 1.25$ (cm)^2/sec and $D_{nB} = 20$ (cm)^2/sec. For homogenous devices we assume Band-gap to be uniform everywhere hence

$E_{gE} = E_{gB} = 1.12$ eV.

Therefore $\beta_F = 1.6 \times 10^4$;

As is evident from Equation XVIII, to obtain high gain, Emitter needs to be heavily doped and we need thin Base Region.

Due to heavy doping in Emitter, degeneracy is introduced which leads to Band Gap Narrowing (BGN) by ΔE_g . The radius of the fifth electron or the donor electron orbiting donor atom is 13×10^{-8}cm = 13Angstrom = 13Aº.

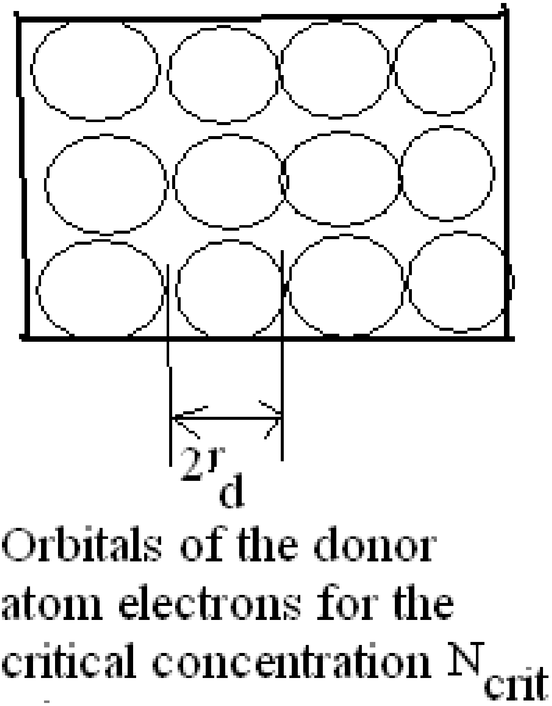

$2r_d$
Orbitals of the donor atom electrons for the critical concentration N_{crit}

Figure VII. Calculation of critical packing density where donor electrons wave function of

donor atoms start to overlap.

$N_{crit} = 1/V_d = 1/(2r_d)^3 = 1.7 \times 10^{19}$/cc.

When the doping level approaches this critical level, energy level corresponding to donor atoms no more remain DISCRETE but they become a continuum in effect extending the conduction band and narrowing the Forbidden Energy Band Gap which is known as Band Gap Narrowing(BGN). This is a degenerate semiconductor and the Fermi-level lies at the edge of the conduction band or within the conduction band as it is in metal. Hence a degenerate semiconductor is semi-metal. The energy band diagram of degenerate and non-degenerate semiconductor is given in Figure VIII.

Heavy doping of emitter causes Band Gap Narrowing in Emitter as shown in Figure VIII.

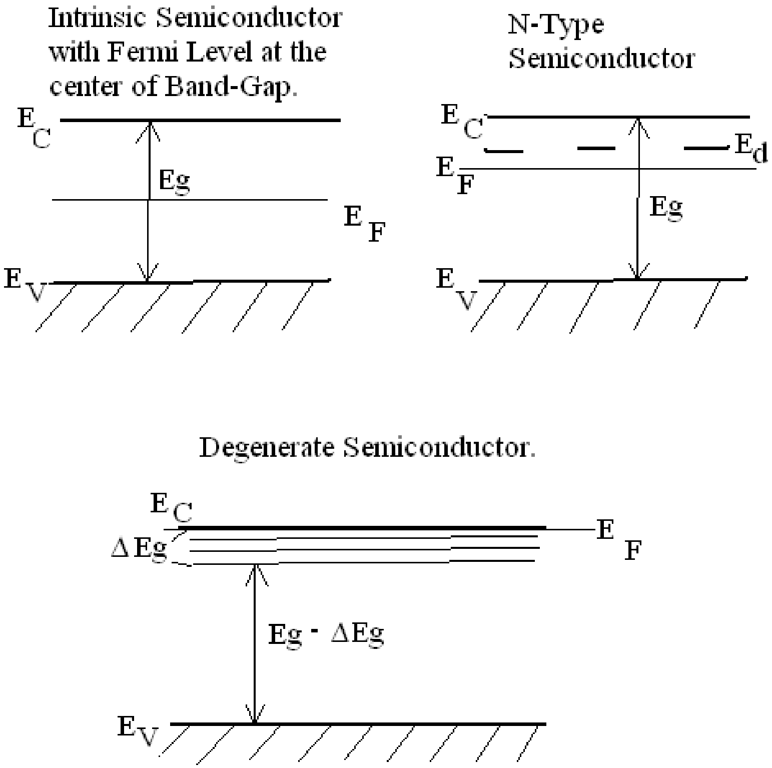

Intrinsic Semiconductor with Fermi Level at the center of Band-Gap.
E_C
Eg
E_F
E_V
N-Type Semiconductor
E_C
E_d
E_F
Eg
E_V
Degenerate Semiconductor.
E_C
ΔEg
E_F
Eg - ΔEg
E_V

Figure VIII. The Energy Band Diagram of Pure, doped and heavily doped semiconductor.

There is Band Gap Narrowing in Emitter and Band Gap in Base remains E_{g0} therefore we get the following expressions

$$\frac{n_{iE}^2}{1} = \frac{N_C N_V Exp(\frac{-(E_{go} - \Delta E_g)}{kT})}{1}$$

$$\frac{n_{iB}^{2}}{1} = \frac{N_C N_V Exp(\frac{-(E_{go})}{kT})}{1}$$

$$\frac{(n_{iE}^{2})}{1} = \frac{n_{iB}^{2} Exp(\frac{\Delta E_g}{kT})}{1} \qquad XIX$$

Substituting Equation XIX in Equation XVIII we get the DC current gain as:

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{1}{N_{A|B}}\right)\left(\frac{N_{D|E}}{Exp(\frac{\Delta E_g}{kT})}\right)$$

XX

Therefore

$$N_{Deff} = N_D Exp(-\frac{\Delta E_g}{kT})$$

XXI

The empirical relationship for BGN is:

$$\Delta E_g = E_{ref}\left\{ln\left(\frac{N_D}{N_{ref}}\right) + \left[\ln\left(\frac{N_D}{N_{ref}}\right)^2 + 0.5\right]^{0.5}\right\}$$

XXII

Where $E_{ref} = 0.009$eV and $N_{ref} = 10^{17}$/cc for typical cases.

Table I. ΔE_g vs N_D .

Emitter Doping Density N_D (number per cc)	BGN ΔE_g (eV)
10^{20}	0.12466
10^{19}	0.083
10^{18}	0.0424
10^{17}	6.36×10^{-3}
10^{16}	1meV

Table II. N_{Deff} vs N_D .

N_D (number per cc)	ΔE_g (eV)	N_{Deff} (number per cc)
10^{20}	0.12466	8.258×10^{17}
10^{19}	0.083	4×10^{17}
10^{18}	0.0424	1.96×10^{17}

10^{17}	6.36×10^{-3}	8×10^{16}
10^{16}	1meV	1×10^{16}

Heavy doping of emitter causes Band Gap Narrowing in Emitter which limits the improvement in Injection Efficiency due to heavy doping of Emitter. Emitter doping may be $N_D = 10^{20}$/cc but in effect we get the Injection Efficiency corresponding to 10^{18}/cc.

For a doping level of $N_D = 10^{20}$/cc, $\Delta E_g = 0.12$eV.

This gives an effective doping level of $N_{Deff} = 10^{18}$/cc.

If we take into account of BGN, we must use Eq.XX .

Using Eq.XX we get : $\beta_F = 160$.

Section III.3. Polysilicon emitter BJT (PEBJT)

The carrier concentration profile in CE BJT in forward active mode is shown in Figure VI. In 60s EB junction was like a wide base diode but with dimension scaling and improvement in technology, EB junction became shallow and hence narrow base diode. In a narrow base diode the emitter surface contact is decisive in determining the resultant emitter injection efficiency. As shown in Figure IX, metal contact directly to pure single crystal emitter portion causes a poorer injection efficiency as compared to that where metal contact and n+ pure Silicon has heavily doped poly-Silicon Emitter sandwiched between the two layers.

***SORRY,
THIS
MEDIA
TYPE IS
NOT
SUPPORTED.***

Figure IX. Linear Gradient is a strong function of surface condition. Metal contact gives a large gradient and Poly Silicon contact to Emitter gives a much lower linear gradient on the emitter side thereby improving I Dn /I D (Injection Efficiency) by several orders of magnitude.

To achieve higher degree of integration we had to go for smaller feature size as well as shallow devices. Metallic contact gives infinite surface recombination velocity, hence it gives a much steeper gradient as shown in Fig. IX a. whereas Poly-Silicon gives a much lower gradient resulting in a very low hole component of the total current thereby giving a much higher injection efficiency.

Table III. Room Temperature current gain as a function of emitter contact for device run BIP-8. ["Effect of Emitter Contact on Current Gain of Silicon Bipolar Devices", T.H.Ning & R.D. Isaac, *IEEE Int. Electron Devices Meeting,* **pp.473-476, 1979].**

Wafer	H	H	E
Contact	Al	Pd_2Si + Al	Poly + Al
R_{DB} (KΩ/□)	7.2 ± 0.2	7.2 ± 0.2	7.6 ± 1.0
Gain	67 ± 3	51 ± 8	158 ± 12

" The current gain of silicon bipolar transistors with shallow emitters depends critically on the emitter contact technology……………The conventional contact by metal or metal silicide degrades the current gain, while contact by a thin layer of poly-silicon is effective in improving the current gain" [ibid].

In Polysilicon emitter BJT (PEBJT), a polycrystalline Si Film lies between Metallic Emitter Contact and Silicon Proper. By ion-implantation, the Polycrystalline Silicon Film is doped with Donor

Arsenic Ions. By subsequent annealing, shallow diffusion of Donor Arsenic atoms from N+ Poly-Silicon Source takes place leading to a shallow layer of N+ in Si proper contacted by a similarly doped N+ Poly-Silicon thin film. This structure is known as Polysilicon emitter BJT.

1. This is suitable for shallow EB Junction;
2. Compatible with self-aligned processing. This minimizes parasitic Resistance and Capacitances leading to faster response time.
3. Short Circuit Current Gain (β_F) in PET is considerably improved as compared to that of shallow all crystalline emitter of equivalent thickness.

III.3.1. The physics of improvement of Short Circuit Current Gain (β_F) in PET.

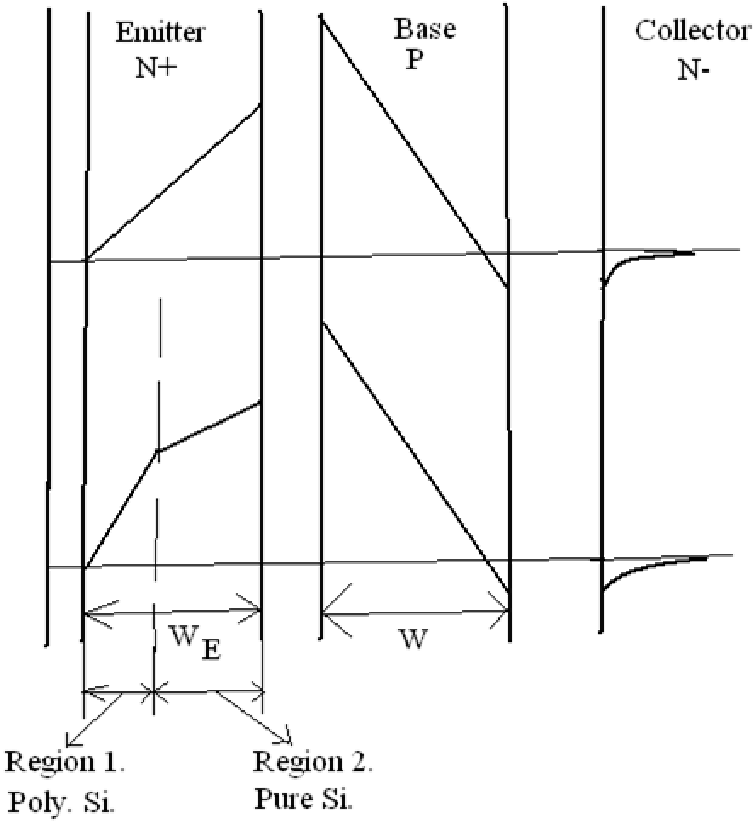


Figure X. Minority Carrier Concentration in a shallow EB Junction case 1. All pure single crystal structure and case 2 in Poly Si Emitter BJT.

Since $W_E << L_{pE}$, therefore there is negligible recombination through out the Emitter. At the interface of Poly Si(region 1) and Pure Si(Region 2), continuity of carrier concentration and continuity of diffusion current requires that:

$$qDe1\frac{dpE1}{dx}$$

=

$$qDe2\frac{dpE2}{dx}$$

$$\frac{dpE1}{dx} = \frac{De2}{De1} \times \frac{dpE2}{dpE1} = \frac{\mu e2}{\mu e1} \times \frac{dpE2}{dx}$$

Therefore carrier concentration gradient in poly Si region is several times the carrier concentration gradient in Pure Silicon Emitter because mobility of hole in Pure Single Crystal Silicon region is several times greater than that in Poly Si on account of domain scattering in Poly region. Hence hole current crossing EB junction into Emitter is considerably reduced leading to a near unity Injection Efficiency. This leads to a much better α and hence a much better β . Now doping of base can be increased leading to a reduced base spreading resistance and hence better Figure of Merit and better Frequency Response without losing on current gain.

Section III.4. Base Spreading Resistance r x and the frequency response of BJT:

Base spreading resistance is due to the narrow sandwich of high sheet resistance P base. P Base is of 200 $\Omega/\square$ to keep $N_{A|B}/N_{D|E} = 10^{17}/10^{19}$ as low as possible. This results in $r_x = 100\ \Omega$. This creates serious deterioration in high frequency performance of BJT especially in terms of Unity Power Gain Frequency (f_{max}).

It can be shown that (f_{max}) $= [f_T/(8\pi r_x C_\mu)]^{1/2}$

Using the dimensions of BJT used in 70s we get, we get

f_t (Unity gain BW) $= 624$MHz;

$f_{max} = [f_T/(8\pi r_x C_\mu)]^{1/2} = 18.6$GHz;

From this example it is clear that base spreading resistance r_x must be minimized to get best frequency performance but reduction of r_x requires increased base conductivity but increased base conductivity means lower Injection Efficiency because

Injection Efficiency $= \gamma = 1/[1 + \sigma_B W/(\sigma_E L_p)]$

From injection efficiency point of view: $\sigma_B << \sigma_E$;

These are contradictory requirements hence to achieve Short Circuit Current Gain of 100 we cannot allow Base Spreading Resistance to go below 100 ohm.

By using Poly Silicon , γ can be improved several orders of magnitude. This results in β of the order of 10,000. Here base conductivity can be increased to minimize Base Spreading Resistance. Thus with a reasonable β of 100, considerably lower Base Spreading Resistance can be achieved which plays crucial role in frequency performance and in improving the Figure of Merit of BJT which happens to be

GBP = $1/(r_x C_\mu)$;

Section III.5. Method of improving Emitter Injection Efficiency by Bipolar Hetero Structure.

Rewriting Equation XVIII we get:

$$\beta_F = \left(\frac{W_E}{W_B}\right)\left(\frac{D_{nB}}{D_{pE}}\right)\left(\frac{N_{D|E}}{N_{A|B}}\right) Exp\left[\frac{(E_{gE}-E_{gB})}{kT}\right]$$

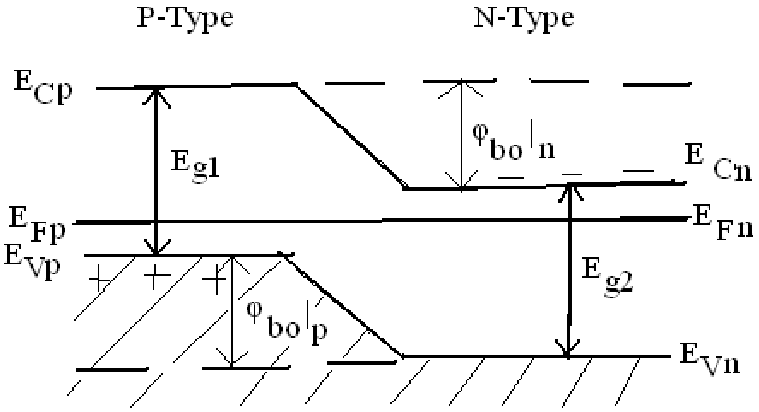
XVIII

So if the EB junction is not a homojunction but a hetro junction

where $(E_{g|E} - E_{g|B}) = 0.15$eV then $\beta_F = 400$ even if $(N_{D|E}/N_{A|E}) = 1$ that is both E and B are equally doped. But $(N_{D|E}/N_{A|E}) = 1$ will drastically reduce Base Spreading Resistance which will leads to improved frequency response as well as improved GBP.

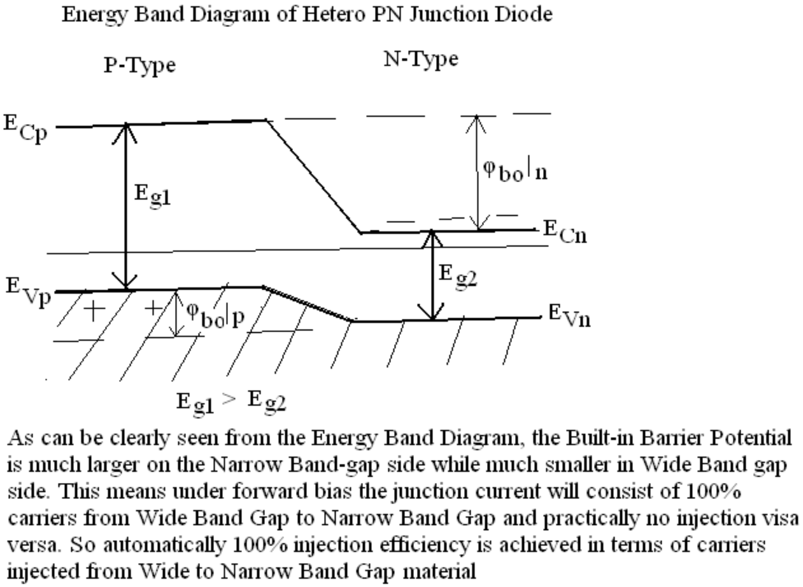
Graphically the improvement is evident as shown in the Graph:

Energy Band Diagram of a HOMO PN JUNCTION.



As can be seen in a HOMO JUNCTION, the built-in barrier potentials (are of equal magnitude on the two sides of the junction hence it is equally difficult for the minority carrier to cross over the potential barrier at the junction under forward bias. $E_{g1} = E_{g2}$

Figure XI. In a HOMO JUNCTION of equal doping on two sides the total forward junction current is made up of 50% electrons being injected from N to P side and 50% holes being injected from P side to N side.



Energy Band Diagram of Hetero PN Junction Diode

As can be clearly seen from the Energy Band Diagram, the Built-in Barrier Potential is much larger on the Narrow Band-gap side while much smaller in Wide Band gap side. This means under forward bias the junction current will consist of 100% carriers from Wide Band Gap to Narrow Band Gap and practically no injection visa versa. So automatically 100% injection efficiency is achieved in terms of carriers injected from Wide to Narrow Band Gap material

Figure XII. In a HETERO JUNCTION under forward bias the Built-in Barrier Potential are so unequal that Junction Current is wholly constituted of carriers injected from Wide Band Gap to Narrow Band Gap thereby automatically achieving 100% Injection Efficiency even without creating large doping differential on the two sides.

Thus we see that heterojunction are very naturally disposed towards giving 100% injection efficiency even with a very low base spreading resistance.

Section III.6. Fermi-level pinning and its prevention.

(academic.brooklyn.cuny.edu/physics/tung/Schottky/index.htm)

At Metal-ntype-Semiconductor interface we realize Schottky Barrier Diode. This exhibits a rectifying contact characteristic hence it has an in-built Schottky Barrier Height(SBH) as illustrated in Figure IV.1.

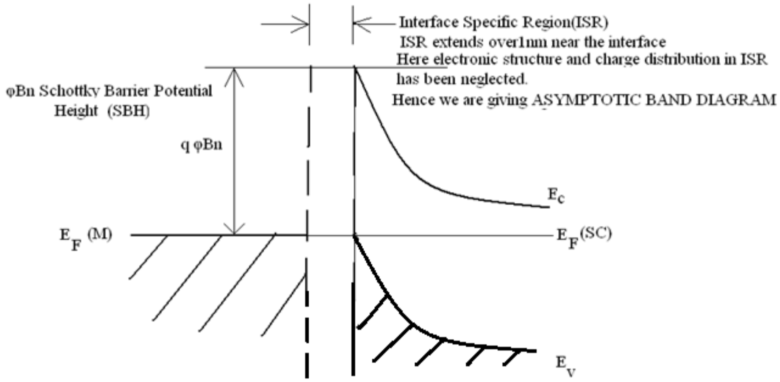

Interface Specific Region(ISR)
ISR extends over1nm near the interface
Here electronic structure and charge distribution in ISR
has been neglected.
Hence we are giving ASYMPTOTIC BAND DIAGRAM
φBn Schottky Barrier Potential
Height (SBH)
q φBn
E_F (M)
E_c
E_F (SC)
E_v

Figure XIII.. Asymptotic Energy Band Diagram of Metal-Semiconductor Interface and illustration of Schottky Barrier Potential Height.

It was found experimentally that SBH is insensitive to the Metal Work Function. This was defined as Fermi-Level pinning (FL pinning). It generally occurred at the mid-band gap of Semiconductor. This was always the case with polycrystalline MS interfaces and it was quite counterintuitive. In 1980s a few high quality, single crystal MS interfaces prepared and SBH measured. SBH was found to be sensitive to orientation/structure of MS interface. By spatially-resolved SBH measurement technique notably by Ballistic Electron Emission

Microscopy(BEEM) it was found that SBH are inhomogeneous at polycrystalline MS interfaces and structure dependent at single crystal MS interfaces.

At the turn of the century when dipole associated with chemical bonding at MS interfaces was modeled using established methods borrowed from Molecular Physics it was shown that FL pinning was a natural consequence of interfacial bonding.

If minimization of total energy is applied

1. then it results in structure dependence of SBH at epitaxial interfaces and
2. SBH in homogeneity at polycrystalline interfaces.

By this theory FL pinning at polycrystalline MS interfaces and the pinning position at mid-band gap comes automatically.

This model is being further refined.

Section III.7. High-k solution for ULSI CMOS. [The high-k solution by Mark T. Bohr, Robert S. Chau, Tahir Ghani & Kaizad Mistry, *IEEE Spectrum*, October 2007, pp 23-29]

As the level of integration evolved, size of CMOS halved every 24 months. At this rate of scaling, by 1998 SiO_2 gate insulation became 5 atomic layer thick with total thickness scaled to 13Aº each atomic layer being 2.6Aº. Current leakage and heating became a serious problem. Because of wave nature of electron and its quantum mechanical tunneling property, thin gate oxides allow the electrons accumulated on gate to leak to the channel. To overcome this problem we need physically thick gate oxides to prevent quantum mechanical tunneling but at the same time electrically thin so that channel is turned at low threshold. If dielectric constant is doubled then thickness can be doubled without any reduction in Turn-ON capability since $C = k\varepsilon_0 A/d$

$$C = \frac{k\varepsilon_0 A}{d}$$

where k = dielectric constant or relative permittivity of the gate insulator

d = thickness of the gate insulator and A = cross-sectional area of the gate insulator.

Metal interconnects are preferred over low-k material so that propagation delay is minimized.

$$v_{propagation} = \frac{1}{\sqrt{\mu_0 k\varepsilon_0}}$$

Low k will give higher velocity of propagation thereby minimizing the propagation delay.

MOS capacitance had to be reproducible, its behavior should be repeatable. To achieve this reproducibility and repeatability, Reactive Sputtering or Metal Organic CVD had to be abandoned in favour of Atomic Layer Deposition . This gave the necessary smoothness to the surface of the electrode as well as precise control over the thickness of the Gate Insulator. This ensured reproducibility and repeatability.

Several dielectrics were studied such as Al_2O_2 , TiO_2 , Ti_2O_2 , Ti_2O_5 , HfO_2 , $HfSiO_4$, ZrO_2, $ZrSiO_2$, La_2O_3 .

With scaling due to Fermi-Level pinning , higher threshold Voltage (V_{Thres}) resulted. Also High-k material have high elasticity hence result in higher phonon scattering or lattice scattering resulting in lower channel mobility. By screening the phonon effect, the deterioration in channel conductivity could be prevented. This required increasing the electron density in Poly-Si gate. Hence we had to switch to Metal-High k combination. This prevented:

1. Fermi-Level pinning hence threshold voltage normalized;
2. This normalized channel conductivity which had drastically deteriorated due to dipole vibration effect;

3. Metal Gate Electrode provides better bonding between Metal-Dielectric. Hence stable operation.

In mid-2003, Intel's Hillsbaro, Ore, Development Fab developed HK-MG CMOS. Intel's 130nm technology was used. Using Hafnium-Based Oxide and Metal Gate electrode following characteristics were achieved:

1. Low Threshold = 1V;
2. Negligible leakage current through Oxide;
3. High 2-D channel mobility;

The standard fabrication method was "Gate First". In this method:

Gate Dielectric + Gate Electrode were laid first;

Source and Drain implanted;

Silicon is annealed to repair the damages that occurred during ion-implantation. High Temperature became a problem for the new technology of HK-MG. So "Gate Last" technology was adopted which circumvented the annealing problem. This led to a paradigm shift in late 2004.

The new flow process was 45nm technology with High k + Metal Gate using Gate-Last strategy. Using this flow process in January 2006, 153Mb SRAM with 1 billion CMOS was built. Leakage gate current was reduced by a factor of 10. But there was sub-threshold leakage.

Scaling had reduced Threshold Voltage but reduced V_{Th} led to increased sub-threshold leakage which defeated the nanoWatt-logic objective. Each new generation of scaled down transistor would increase I_{ON} by about 30% but would lead to $I_{sub\text{-}threshold}$ increase by 3-Fold. So at ULSI level, power efficiency and low leakage would be at premium rather than speed.

Table IV.The options at 45nm HK-MG CMOS Technology

	Option 1	Option 2
Oxide thickness	13Aº	26Aº
I_{ON}	25% increase	No increase
$I_{sub\text{-}threshold}$	No increase	1/5 I_{sub}
V_{Thres}	Same as before	Increased threshold

The CMOS circuits are built between these two extremes.

PENRYN dual-core µP has 410mCMOS and PENRYN quad-core μP has 820 m CMOS. These are optimized for mobile applications, desk-top computers, 64-bit workstations and server applications.

Section III.8. Atomic layer Deposition

Dielectric Layers were deposited by Reactive Sputtering or by Metal Organic CVD. This left unevenness on the surface which caused charge trapping. There were charges stored at the interface too. These charges altered the behavior of MOS capacitance from discharge cycle to discharge cycle. So a deposition method had to be adopted which allowed dielectric deposition in controlled manner atom layer by layer.

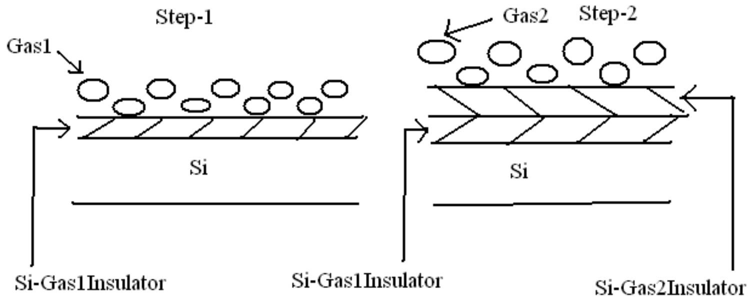


Figure XIV. Methodology of Atomic Layer Deposition.

In Figure XIV the steps taken for Atomic Layer Deposition is illustrated. Gas1 reacts with bare Silicon surface to form a single atomic layer of insulator Si-Gas1. As soon as one layer is completed

Gas1 stops reacting with Si because no bare Si is in contact with Gas1.

In step 2, Gas2 is chosen which is reactive with dielectric1 or Si-Gas1 insulator. Gas2 reacts with dielectric layer1 to form dielectric layer2 (Dielectric1-Gas2 insulator). As soon as dielectric layer2 covers the whole of dielectric1, reaction stops. Thus each reaction is terminated at the end of layer deposition. In this way we achieve layered gate insulator which is controllable down to the width of a single atom.

This produces much smoother dielectric hence the charge trapping is prevented . MOS produced is reproducible and stable in operation.

SSPD_Chapter 6_Part 9_Figure IX of Appendix on Device Physics

SSPD_Chapter 6_Part 9_Figure IX of the Appendix on Device Physics is being uploaded.

SSPD_Chapter 6_Part 9_Figure IX of the Appendix on Device Physics

Section III.3. Polysilicon emitter BJT (PEBJT)

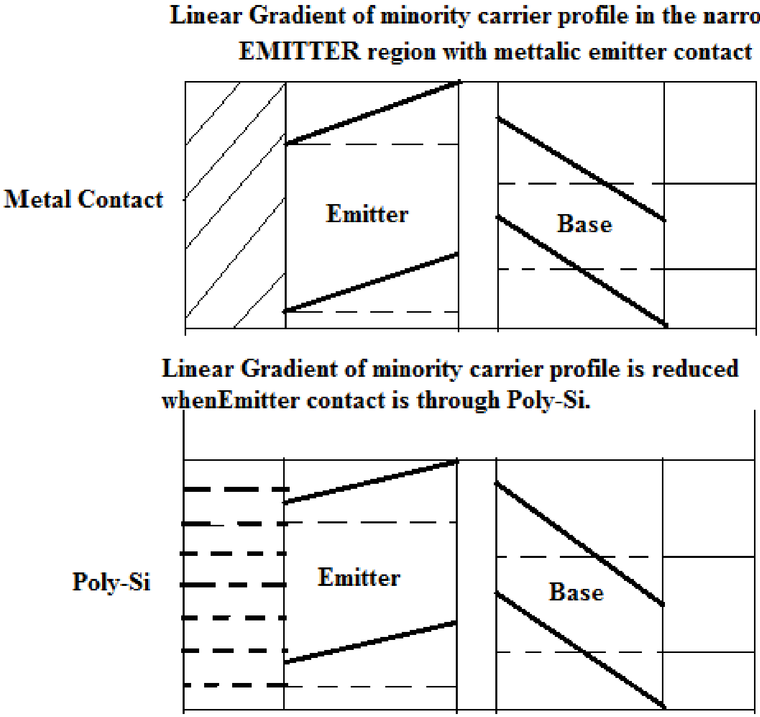

Linear Gradient of minority carrier profile in the narrow EMITTER region with mettalic emitter contact
Metal Contact
Emitter
Base
Linear Gradient of minority carrier profile is reduced whenEmitter contact is through Poly-Si.
Poly-Si
Emitter
Base

Figure IX. Linear Gradient is a strong function of surface condition. Metal contact gives a large gradient and Poly Silicon contact to Emitter

gives a much lower linear gradient on the emitter side thereby improving I Dn /I D (Injection Efficiency) by several orders of magnitude.

SSPD_Chapter 6_Part 10_Simulation of MOSFET
SSPD_Chapter 6_Part 10 works out the example number one on MOSFET.

SSPD_Chapter 6_Part 10_MOSFET Simulation.

6.10.1. mos1ex01:Plot of I D -V GS at constant V DS and extraction of V Th .

```
go athena

#

line x loc=0.0 spac=0.1

line x loc=0.2 spac=0.006

line x loc=0.4 spac=0.006

line x loc=0.6 spac=0.01

#

line y loc=0.0 spac=0.002

line y loc=0.2 spac=0.005

line y loc=0.5 spac=0.05

line y loc=0.8 spac=0.15
```

```
#

init orientation=100 c.phos=1e14 space.mul=2

#pwell formation including masking off of the nwell

#

diffus time=30 temp=1000 dryo2 press=1.00 hcl=3

#

etch oxide thick=0.02

#

#P-well Implant

#

implant boron dose=8e12 energy=100 pears

#

diffus temp=950 time=100 weto2 hcl=3

#

#N-well implant not shown -

#
```

```
# welldrive starts here

diffus time=50 temp=1000 t.rate=4.000 dryo2 press=0.10 hcl=3

#

diffus time=220 temp=1200 nitro press=1

#

diffus time=90 temp=1200 t.rate=-4.444 nitro press=1

#

etch oxide all

#

#sacrificial "cleaning" oxide

diffus time=20 temp=1000 dryo2 press=1 hcl=3

#

etch oxide all

#

#gate oxide grown here:-
```

```
diffus time=11 temp=925 dryo2 press=1.00
hcl=3

#

# Extract a design parameter

extract name="gateox" thickness oxide
mat.occno=1 x.val=0.05

#

#vt adjust implant

implant boron dose=9.5e11 energy=10 pearson

#

depo poly thick=0.2 divi=10

#

#from now on the situation is 2-D

#

etch poly left p1.x=0.35

#

method fermi compress
```

```
diffuse time=3 temp=900 weto2 press=1.0
#
implant phosphor dose=3.0e13 energy=20 pearson
#
depo oxide thick=0.120 divisions=8
#
etch oxide dry thick=0.120
#
implant arsenic dose=5.0e15 energy=50 pearson
#
method fermi compress
diffuse time=1 temp=900 nitro press=1.0
#
# pattern s/d contact metal
etch oxide left p1.x=0.2
deposit alumin thick=0.03 divi=2
```

```
etch alumin right p1.x=0.18

# Extract design parameters

# extract final S/D Xj

extract name="nxj" xj silicon mat.occno=1
x.val=0.1 junc.occno=1

# extract the N++ regions sheet resistance

extract name="n++ sheet rho" sheet.res
material="Silicon" mat.occno=1 x.val=0.05
region.occno=1

# extract the sheet rho under the spacer, of the LDD
region

extract name="ldd sheet rho" sheet.res
material="Silicon" \

mat.occno=1 x.val=0.3 region.occno=1

# extract the surface conc under the channel.

extract name="chan surf conc" surf.conc
impurity="Net Doping" \

material="Silicon" mat.occno=1 x.val=0.45

# extract a curve of conductance versus bias.
```

```
extract start material="Polysilicon" mat.occno=1 \

bias=0.0 bias.step=0.2 bias.stop=2 x.val=0.45

extract done name="sheet cond v bias" \

curve(bias,1dn.conduct material="Silicon"
mat.occno=1 region.occno=1)\

outfile="extract.dat"

# extract the long chan Vt

extract name="n1dvt" 1dvt ntype vb=0.0
qss=1e10 x.val=0.49

structure mirror right

electrode name=gate x=0.5 y=0.1

electrode name=source x=0.1

electrode name=drain x=1.1

electrode name=substrate backside

structure outfile=mos1ex01_0.str

# plot the structure

tonyplot mos1ex01_0.str -set mos1ex01_0.set
```

```
############# Vt Test : Returns Vt, Beta and Theta ################

go atlas

# set material models

models cvt srh print

contact name=gate n.poly

interface qf=3e10

method newton

solve init

# Bias the drain

solve vdrain=0.1

# Ramp the gate

log outf=mos1ex01_1.log master

solve vgate=0 vstep=0.25 vfinal=3.0 name=gate

save outf=mos1ex01_1.str

# plot results

tonyplot mos1ex01_1.log -set mos1ex01_1_log.set
```

```
# extract device parameters

extract name="nvt"
(xintercept(maxslope(curve(abs(v."gate"),abs(i."drain"))))
\

- abs(ave(v."drain"))/2.0)

extract name="nbeta"
slope(maxslope(curve(abs(v."gate"),abs(i."drain"))))
\

* (1.0/abs(ave(v."drain")))

extract name="ntheta" ((max(abs(v."drain")) *
$"nbeta")/max(abs(i."drain"))) \

- (1.0 / (max(abs(v."gate")) - ($"nvt")))

Quit
```

We have to create the initial structure file.

By Mesh Define Menu, initial rectangular grid is specified. By selecting the VIEW……. Button, we can click the View Grid Window. We can adjust the grid lines according to our requirement. In critical areas we want fine mesh and in passive, non-critical area we require coarse mesh. Accordinfgly we set the grid line and write it into the INPUT file. It comes as shown below:

```
#
line x loc=0.0 spac=0.1
line x loc=0.2 spac=0.006
line x loc=0.4 spac=0.006
line x loc=0.6 spac=0.01
#
line y loc=0.0 spac=0.002
line y loc=0.2 spac=0.005
line y loc=0.5 spac=0.05
line y loc=0.8 spac=0.15
#
```

These lines describe the GRID or MESH

Now we define the Initial Substrate. By Mesh Initialize…….Mesh Initialize Menu appears.

Here we have three options for ‘orientation’. For BJT we will select <111> but for MOS structures we will select <100>. This helps minimize the interface impurities and oxide trapped impurities. This in turn helps ensure a stable device.

Background doping is selected by clicking Phosphorous Box. Its concentration is set at 1 by the slider and 14 is written into exponent box.

Dimensionality is chosen at AUTO. This will start in 1-D and automatically switch to 2-D.

Grid Spacing scaling factor is chosen at 2.

By WRITE button the information is written into the input file. It is as shown below.

```
init orientation=100 c.phos=1e14 space.mul=2
```

Using these two informations (Grid Information and Grid Initialization) if we RUN Athena we get the INITIAL STRUCTURE as Initial Triangular Grid.

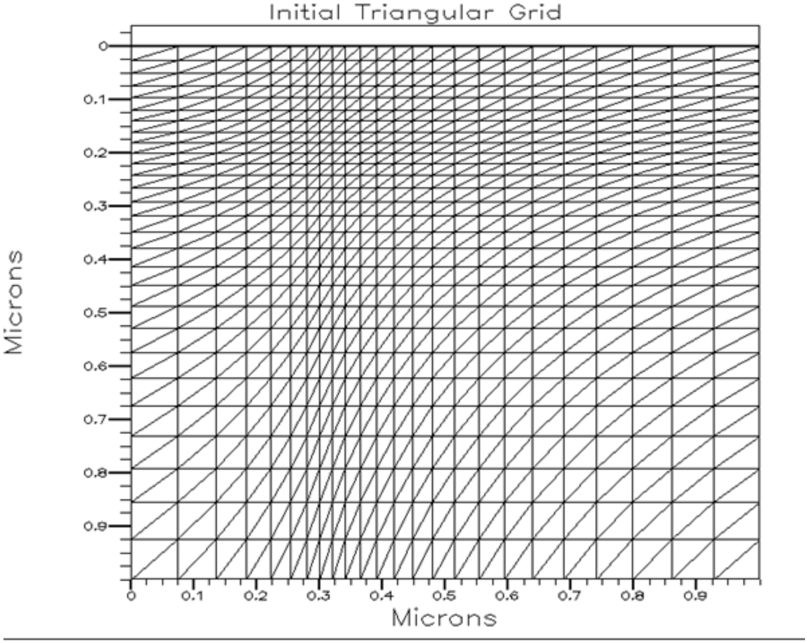


Figure 1. Initial Triangular Grid.

This triangular grid is not drawn according to the specifications given for this example.

The Grid consists of points connected to form a number of triangles. The

physical device is 3D structure but we are showing 2D Grid structure. Hence a point in 2D mesh may correspond to different regions in 3rd dimension. It may belong to Silicon at z = 0µm and to Oxide region at z = 2µm. Hence the same point will have two nodes. A node represents the solution of the coupled equations in the region to which node belongs. Hence in this case a point will have two nodes: one node in Silicon region and another node in Oxide region.

The INIT statement creates $<100>$ Silicon region of $1\mu m \times 1\mu m$ size which has an uniform N-type doping of Phosphorous of doping concentration of 1×10^{14}Phos atoms per cc. This simulation structure is ready for any process step e.g. implant, diffusion, reactive ion-etching.

Now structure manipulation statements will follow. These will precede or alternate with PHYSICAL PROCESS STEPS which will be carried out by SUPREME4, ELITE or OPTOLITH modules.

Thermal Process Simulation, such as diffusion, oxidation or silicidation, is the focal point of SUPREME4.

When we open the ATHENA Diffuse Menu, we have the following options:

#Time is to be adjusted from 0 to 500 minutes.
############################

#We have Temperature option which can be chosen from 500°C to 1300°C.####

#The ‘Ambient’ option has Dry O_2, Wet O_2(Argon or Nitrogen is bubbled through water),N_2, Gas Flow.##################################

#We have gas pressure option. We can choose from 1 atmosphere to 10 atmosphere.############################

#HCl percentage option.###############################

#Comment############################

We have the option of ramping up or ramping down the temperature as we do in Rapid Thermal Anneal(RTA).

Small concentration of HCl , in this case 3%, helps reduce Oxide defect densities and reduce contamination levels.

$4HCl + O_2 \rightarrow 2H_2O + 2Cl_2$.

HCl reacts with Oxygen to form Chlorine. Chlorine reacts with trace metals and other contaminants in the furnace to produce volatile chlorides as by products.

We have the following statements.

#pwell formation including masking off of the nwell

This statement means that we will be process simulating the (E)NMOS in P Well only and extracting the threshold voltage from its transfer characteristics.

```
#

diffus time=30 temp=1000 dryo2 press=1.00 hcl=3
```

In the above statement we are doing dry oxidation for 30 minutes at 1000°C in 1 atmospheric pressure with a mixed ambient which contains 3% HCl acid vapour.

The above parameters of dry diffusion gives 0.03µm thick oxide from the look-up table.

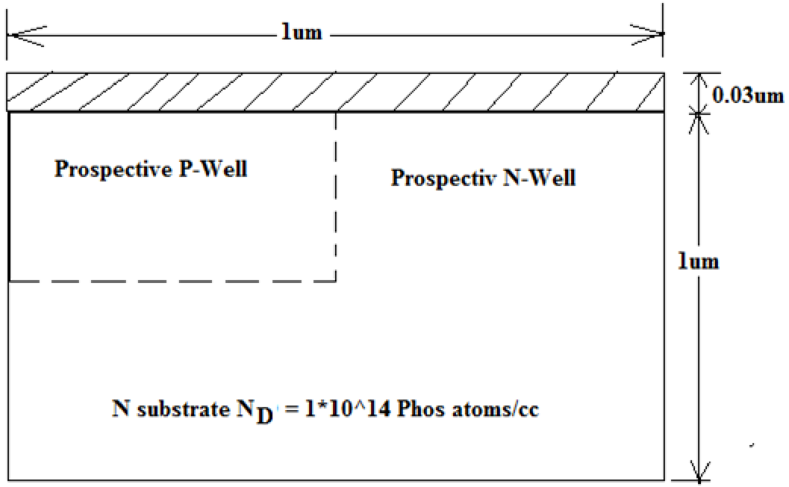


Figure 2. Initial structure with entire surface at y = 0 being covered with 0.03umthick oxide layer by dry oxidation.

The oxide is retained over N Well and completely etched out from P Well portion.

For this we open Athena Etch Menu:

#Etch Method: Geometrical-Machine##############################

#Geometrical Type:All-Left-Right-Dry Thickness-Any Shape############################

#Material- Oxide or Poly Si.##################################

#Etch Location##############################

Geometrical Etch is the default method of Etch.

We can choose the material to be etched and we can specify the portion to be etched.

We may etch out the entire surface(All) or RIGHT of a point situated at xo from the left edge or LEFT of a point situated at x1 from the right edge or the entire thickness over the entire surface area or etch out in a specific polygonal area as described by four or more points in the menu.

#

```
etch oxide thick=0.02
```

Since our previous step was Oxidation 0.03μm thick hence in P-Well region again 0.02 μm has to be etched out leaving 10nm screen oxide over P-Well which will randomize the Boron implant into P-Well region. 30nm thick oxide will completely shield the N-Well region from Boron Implant..

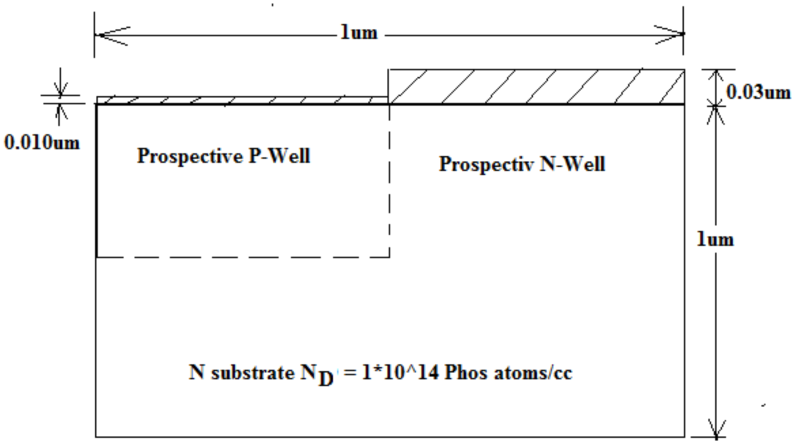


Figure 3. 0.02um thick oxide layer is etched out from P-Well region. In N-Well region entire thickness of Oxide is left intact.

Now we have to carry out the Boron Implant for P-Well.

Ion-implantation is the main method to introduce doping impurities into semiconductor structure. This allows critical dimensions , shallow implant depth, high doses and tilted implants.

We select Process → Implant….. in command menu and open the Implant Menu.

#Impurities: B, P, As, Bf2, Sb, Si, Zn, Se, Be, Mg, Al, Ga, C###############################

#Dose (ions/

```
cm^2):###############################

#Energy
(KeV);###############################

#Model: Dual Pearson-Gauss-Predict-
Montecarlo############################

#Tilt:################################

#Rotaion:#############################

#Material Type: Crystalline-
Amorphous;############################

#Point Defects: None-Unit Damage-
Damage;##############################

#Comment:############################
```

After setting the parameters and the models for Boron Implant in P-well we get the following statement:

```
#P-well Implant

#

implant boron dose=8e12 energy=100 pears
```

In P-Well implant we have used 8×10^{12} Boron ions/cm^2 at 100KeV and Dual Pearson Model has been invoked. With these parameters the peak of the implant will occur at 0.25μm.

```
#

diffus temp=950 time=100 weto2 hcl=3
```

Now wet Oxidation done at 950°C for 100 minutes with 3% HCl. This gives 0.5μm thick oxide layer which covers P-Well again

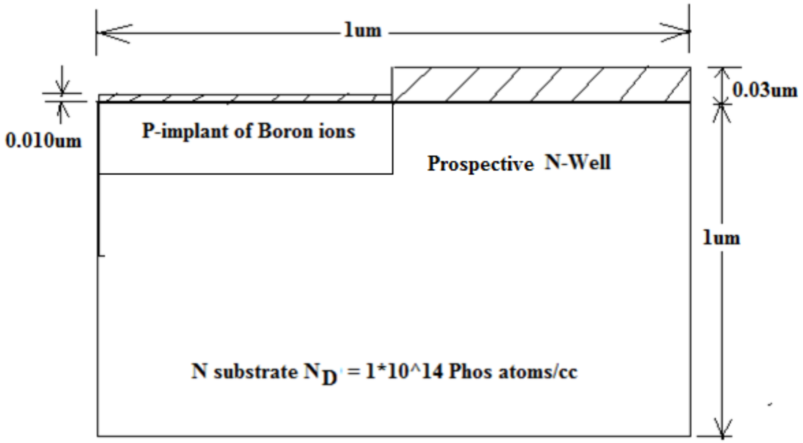


Figure 4. Boron Implant in P-Well region with projected range at ym = 0.5um and peak concentration ym at 0.25um

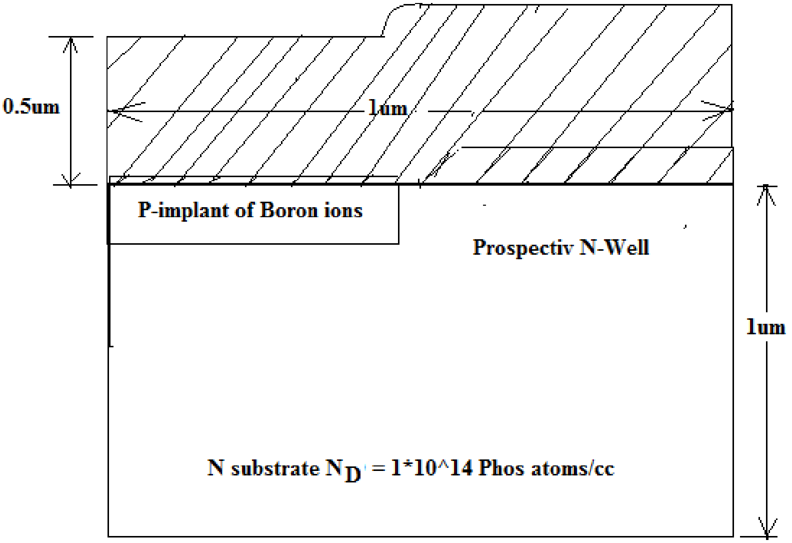


Figure 5. Wet oxidation for covering up the exposed surface of P-Well. Wet oxidation is 0.5um thick.

#

#N-well implant not shown -

#

Now we do thermal annealing to heal the damage as well as to drive the Boron atoms deep into P well.

# welldrive starts here

diffus time=50 temp=1000 t.rate=4.000 dryo2 press=0.10 hcl=3

Drive in takes place at 1000°C for 50 minutes at ramp rate of 4°C per minute with HCl 3% and dry oxygen ambient at 0.1 atmosphere pressure..

```
#

diffus time=220 temp=1200 nitro press=1

#

diffus time=90 temp=1200 t.rate=- 4.444 nitro press=1

#
```

Then damage annealing is done at 1200°C for 220 minutes in Nitrogen Ambient at 1 atmosphere pressure. The goal of damage annealing is is to remove the primary damage created by the implant and restore the silicon lattice to its perfect crystalline state leaving the dopents in active substitutional sites. This annealing removes the damage as well as activates the dopents. Activaion means dopents contribute to electrical conductivity and this can happen only after they occupy substitutional sites.

Next it is ramped down from 1200°C to Room Temperature at a rate of

– 4.444°C/min in Nitrogen Ambient at 1 atmosphere.

```
etch oxide all

#
```

```
#sacrificial "cleaning" oxide
```

Next by geometrical etching all oxide is removed from P-Well.

```
diffus time=20 temp=1000 dryo2 press=1 hcl=3
```

Next 0.03µm Oxide is grown over P-Well at temperature 1000°C in a period of 20 minutes in dry oxygen at a pressure of 1 atmosphere with HCl 3%.

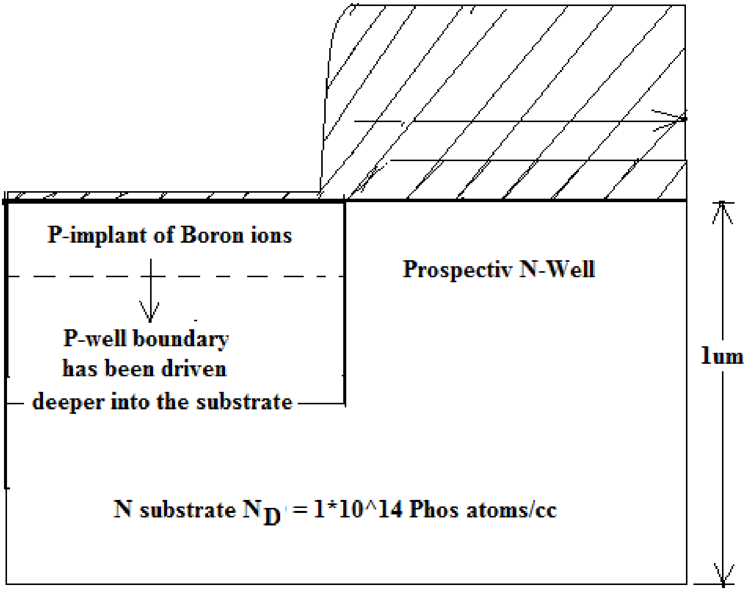


Figure 6. 30nm thick Gate Oxide grown over P-well by dry oxidation.

```
#
```

```
etch oxide all
```

```
#

#gate oxide grown here:-

diffus time=11 temp=925 dryo2 press=1.00 hcl=3

#

# Extract a design parameter

extract name="gateox" thickness oxide mat.occno=1 x.val=0.05
```

A thin oxide layer is grown and again etched. This is done to improve the silicon surface which has received considerable battering due to Boron implants. This is in addition to anneal step.

Again a thin layer of oxide is grown by 11 minutes dry oxidation at 925°C at a pressure of 1 atmosphere and HCl 3%.

The gate oxide is 20nm thick.

```
# Extract a design parameter

extract name="gateox" thickness oxide mat.occno=1 x.val=0.05

#
```

This statement extracts the gate-oxide thickness in the vertical section at xm = 0.05µm. The gate oxide thickness should be 11nm.

```
#vt adjust implant

implant boron dose=9.5e11 energy=10 pearson
```

To precisely control the threshold voltage in the range of 0.5V to 0.8V we do a second ion implantation.

Boron Implant Dose = $9.5\times10_{11}$Boron ions/cm$_2$ and accelerating voltage is 10KeV and we invoke Dual Pearson Model.

The implant dose can be calculated from the formula:

$$V_{Th} = V_{FB} + 2\psi_F + \frac{\sqrt{2\,\epsilon_S\, qN_A(2\psi_F)}}{C_{OX}} + \frac{qQ_I}{C_{OX}}$$

V_{FB} = flat band voltage;

$\Psi_F = (E_i\text{-}E_F)/q$;

Q_I = implanted dose;

$C_{OX} = \varepsilon_{oxide}/t_{oxide}$;

```
#

depo poly thick=0.2 divi=10

#

#from now on the situation is 2-D

#

etch poly left p1.x=0.35

#
```

Here we cover the P-Well with 0.2µm thick Poly_Si as shown in Figure 7.

This Poly_Si is etched left to x = 0.35µm as shown in Figure 8. The N Well is kept out of process simulation hence it is being shown as dotted structure.

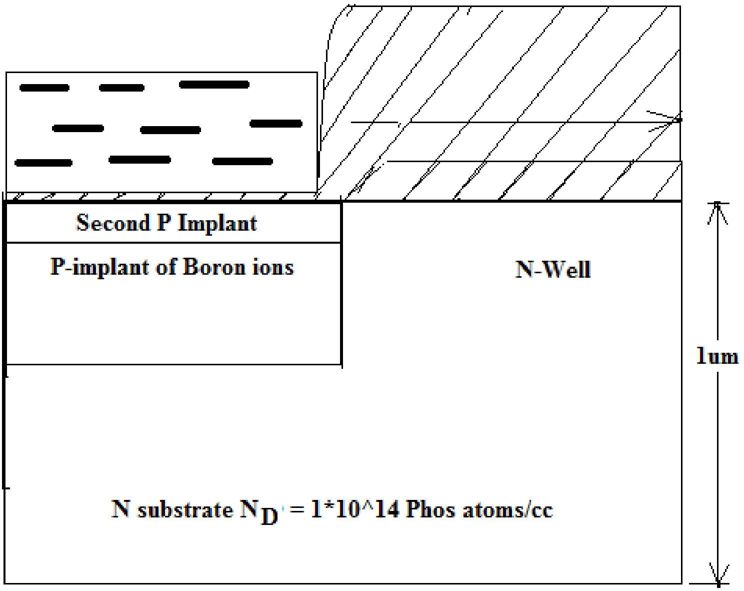


Figure 7. Poly_Si Gate contact deposition 0.2um thick.

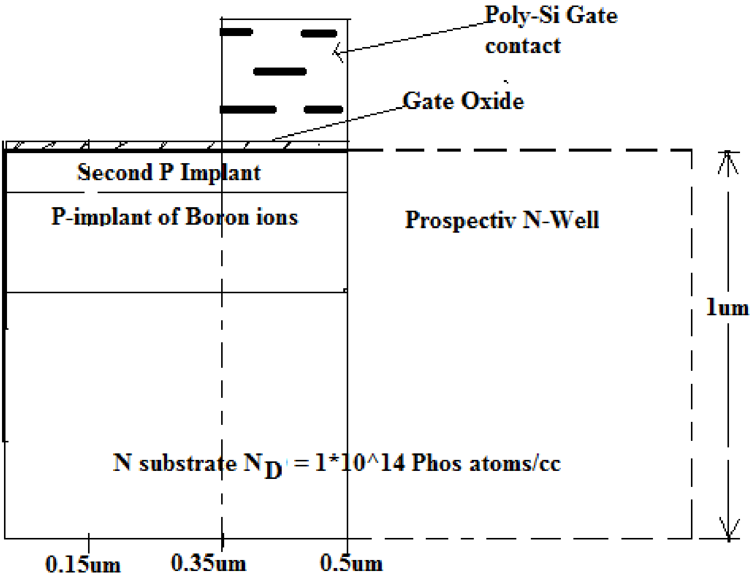


Figure 8. Poly-Si is etched out left to a point xm = 0.35um

```
method fermi compress

diffuse time=3 temp=900 weto2 press=1.0

#
```

Method Fermi Compress is invoked. We are using Fermi model which is applicable to structures having low damage and doping concentration below 1×10^{20} dopents per cc.

3 minutes wet oxidation at 900°C is done at 1 atmosphere.

0.02μm oxide layer is obtained covering the entire surface as shown in Figure 9. It should be noted that this is wet oxidation and not conformal deposition. Hence Oxide layer is formed only in regions where Silicon is available. Necessity of this step is not clear.

```
implant phosphor dose=3.0e13 energy=20 pearson

#
```

Phosphorous implant of Dose 3×10^{13} ions/cm^2 is carried out at an energy of 20KeV.

This phosphorous implant creates self-aligned LDD N-type Source as well as highly N-type doped Poly-Si Gate contact.

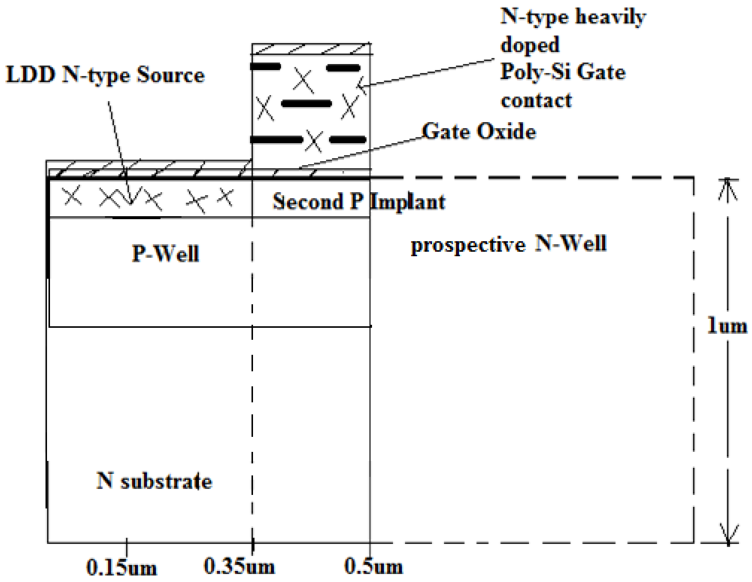


Figure 9. Formation of LDD N-type Source.

Side-Wall Spacer formation can be obtained as follows:

depo oxide thick=0.120 divisions=8

#

etch oxide dry thick=0.120

#

Here we deposit 0.12μm thick oxide and again etch the same amount. Because of conformal nature of deposition during the etching phase side-wall oxide is retained as shown in Figure 10.

Next we do Arsenic Ion-Implant in a larger dose:

implant arsenic dose=5.0e15 energy=50 pearson

#

Arsenic has very little diffusion coefficient hence during subsequent heat cycles it is not redistributed. Plus N+ Source has to be formed hence Arsenic dose is larger and higher accelerating voltage is used to make implant range deeper. Compare this implant with LDD implant. The resulting structure is shown in Figure 11.

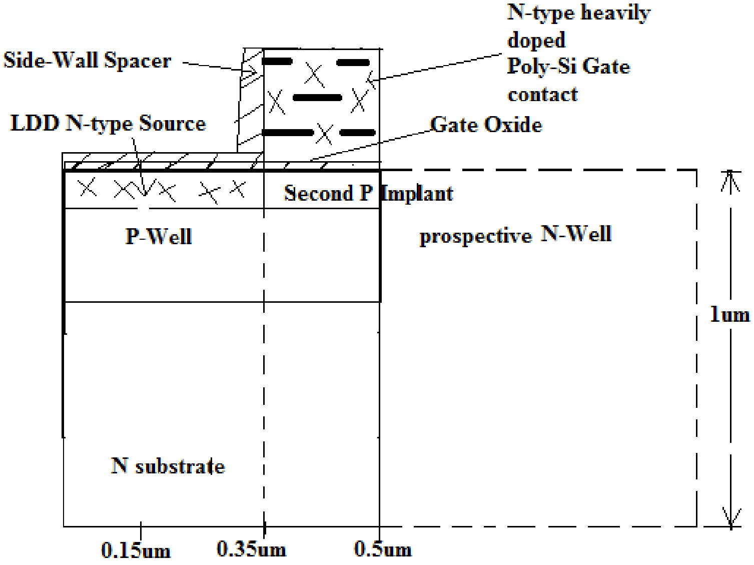


Figure 10. Side-Wall Spacer formation by conformal deposition of 0.12um thick oxide and again etching the same amount.

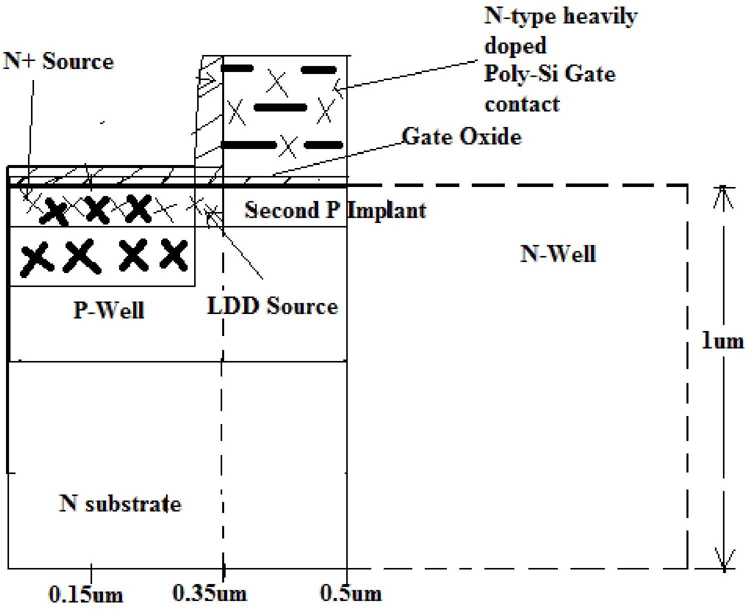


Figure 11. Arsenic implantation in larger dose and deeper penetration to form N+ Source.

Now Metal contact has to be made to N+ source.

```
method fermi compress

diffuse time=1 temp=900 nitro press=1.0

#
```

Rapid annealing is done for 1 minute at 900°C

```
# pattern s/d contact metal

etch oxide left p1.x=0.2

deposit alumin thick=0.03 divi=2
```

Now Oxide is etched out left of the grid line xm=0.2µm.

Next Aluminum is deposited for Source Contact. The entire surface gets covered by Metal Layer shown in Black by conformal deposition method. As shown in Figure 12.

Next we etch out Metal(Al) from the entire surface except the Source contact which is made from xm = 0µm to xm = 0.18 µm. That is Source contact is 0.18µm long extending from the left edge to xm = 0.18 µm. Hence we give the command:

```
etch alumin right p1.x=0.18
```

The resulting structure is shown in Figure 13.

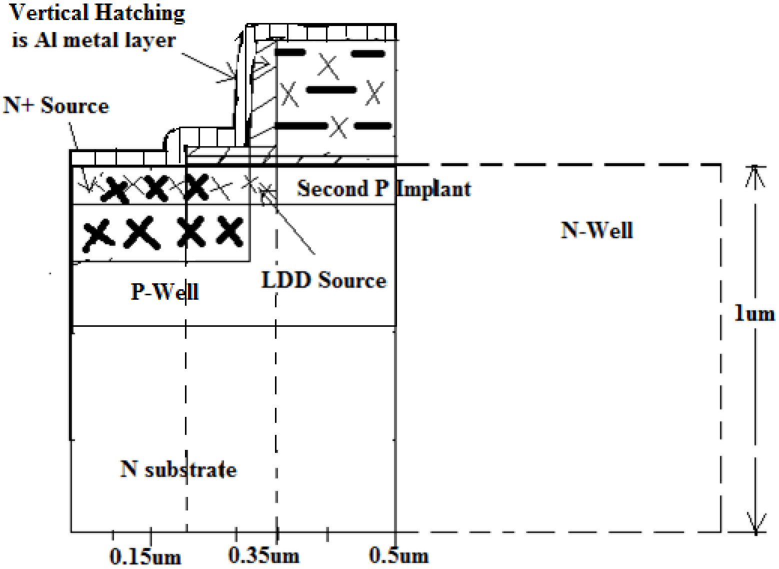


Figure 12. Left of xm= 0.2um, oxide is etched and Aluminum Deposition done. Entire surface gets covered with Aluminum.

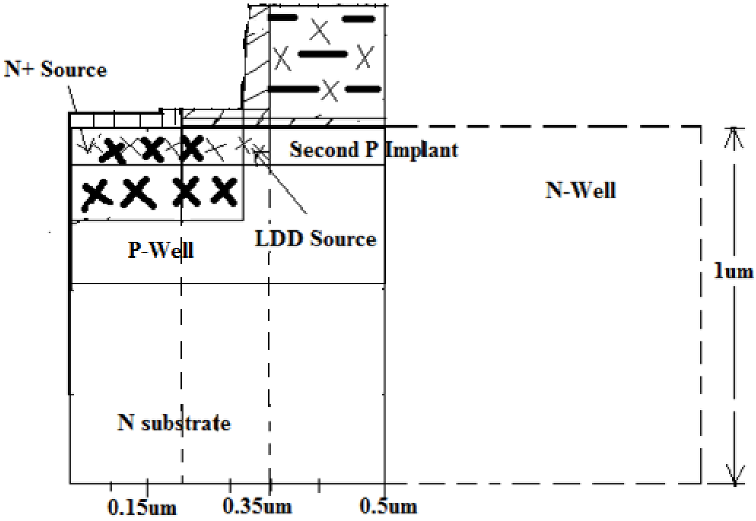


Figure 13. Source Metal Contact is retained and remaining metal right of xm = 0.18um is etched out.

# Extract design parameters

```
# extract final S/D Xj

extract name="nxj" xj silicon mat.occno=1
x.val=0.1 junc.occno=1

# extract the N++ regions sheet resistance

extract name="n++ sheet rho" sheet.res
material="Silicon" mat.occno=1 x.val=0.05
region.occno=1

# extract the sheet rho under the spacer, of the LDD
region

extract name="ldd sheet rho" sheet.res
material="Silicon" \

mat.occno=1 x.val=0.3 region.occno=1

# extract the surface conc under the channel.

extract name="chan surf conc" surf.conc
impurity="Net Doping" \

material="Silicon" mat.occno=1 x.val=0.45
```

In the above statements following fabrication parameters are being extracted at the vertical sections of following physical locations of the MOS structure as shown in Figure 14.

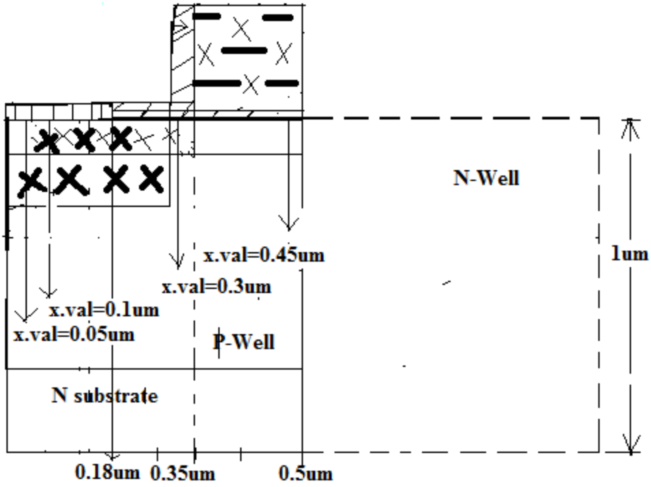


Figure 14. Physical location of the vetical sections at which the fabrication parameters are to be extracted.

At x.val = 0.05μm, Sheet Resistance of N+ + Source region is to be extracted.

At x.val = 0.15μm, Junction depth of Source region is to be extracted.

At x.val = 0.3μm, Sheet Resistance of LDD is to be extracted.

At x.val = 0.05μm, Sheet Resistance of the channel below the gate-oxide is to be extracted.

# extract a curve of conductance versus bias.

extract start material="Polysilicon" mat.occno=1 \

bias=0.0 bias.step=0.2 bias.stop=2 x.val=0.45

```
extract done name="sheet cond v bias" \

curve(bias,1dn.conduct material="Silicon"
mat.occno=1 region.occno=1)\

outfile="extract.dat"
```

Here the Gate Bias voltage is increased from 0V to 2V in step of 0.2V and sheet conductance versus Gate Bias Voltage curve is determined at the vertical section located at xm = 0.45µm.

```
# extract the long chan Vt

extract name="n1dvt" 1dvt ntype vb=0.0
qss=1e10 x.val=0.49
```

In this statement Threshold Voltage is extracted with bias voltage at 0V and Qss = 1×10^{10} at the vertical section situated at xm = 0.45µm.

```
structure mirror right
```

By this statement we are able to generate the whole (E)NMOS structure.

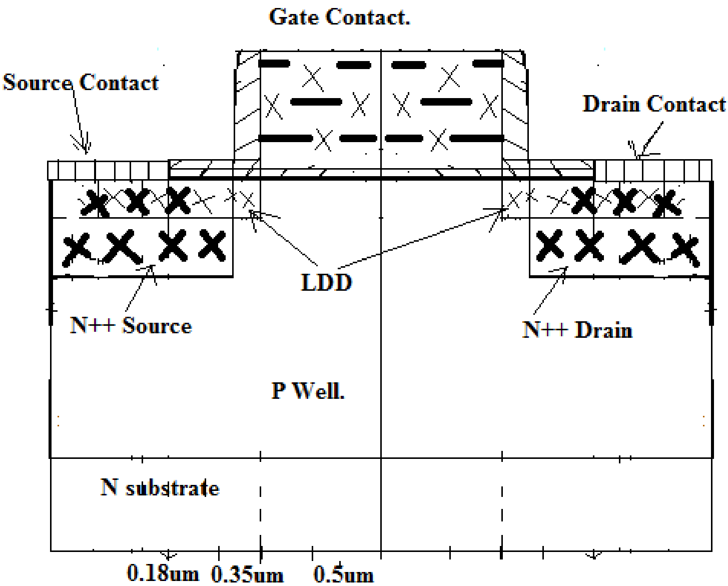


Figure 15. By MIRROR RIGHT command whole structure is generated. Now we have the complete (E)NMOSFET.

```
electrode name=gate x=0.5 y=0.1

electrode name=source x=0.1

electrode name=drain x=1.1

electrode name=substrate backside

structure outfile=mos1ex01_0.str

# plot the structure

tonyplot mos1ex01_0.str -set mos1ex01_0.set

############# Vt Test : Returns Vt, Beta and Theta ################
```

In the last four statements we are giving the name and location of the contact pads of the device generated.

We have Gate Contact Pad located at $xm=0.5\mu m$ and $ym = -0.1\ \mu m$;

Source Contact Pad at $xm = 0.1\ \mu m$ and on the surface of the device;

Drain Contact Pad at $xm = 1.1\ \mu m$ and on the surface of the device;

There is contact pad to the body at the backside.

Beta is the distribution function describing the dopent distribution.

Theta is the distribution function dealing with the interstitials.

SSPD_Chapter 6_Part 10_Simulation of MOSFET_conclusion

SSPD_Chapter 6_Part 10 carries out the device simulation and characterization.

SSPD_Chapter 6_Part 10_MOSFET Simulation2

In MOSFET Simulation1 we used Process Simulator(Athena) to generate the basic (E)NMOS device in P-Well formed in lightly doped N-body. Now we will use ATLAS for device testing and parameterization

```
go atlas
```

```
# set material models
```

```
models cvt srh print
```

```
MODEL MOS PRINT
```

The above statement invokes all the relevant models in context of parameterization of MOS device namely:

CVT , SRH, FERMIDIRAC.

SRH is Shockley-Hall-Reed model deals with indirect recombination through traps and defects.

FERMIDIRAC is fermidirac statistics

$$P(E) = \frac{1}{1 + Exp\,\frac{E - E_F}{kT_L}} \qquad 1$$

CVT is a stand-alone model which accounts for the 2D nature of the inversion layer.

Mobility degradation in 2D inversion layer, which forms the channel, is handled by three different methods:

1. SURFMOB- mobility degradation due to interface scattering (the interface between silicon substrate and gate oxide);
2. SHIRAHATA – transverse electric field model. A transverse electric field exists along the channel;
3. CVT, YAMAGUCHI, TASCH- all these three are standalone model for accounting the mobility degradation in the inversion layer. CVT accounts for transverse electric field effect, doping dependence and temperature dependence.

contact name=gate n.poly

interface qf=3e10

In the above two statements we are naming the Gate Contact and giving the interface statements. The interface statements enable define interface fixed charges, surface recombination velocity and thermionic emission. Here we are enabling a fixed charge of 3×10^{10} at the interface of Si and Oxide.

```
method newton
```

```
solve init
```

This states that we re using coupled Newton Method of numeric calculations. Newton Method is used where we have strongly coupled system with quadratic convergence. But this requires a more accurate initial guess to the problem to obtain convergence.

```
# Bias the drain
```

```
solve vdrain=0.1
```

We are solving for drain current with drain voltage at 0.1V and gate voltage at 0V

```
# Ramp the gate
```

```
log outf=mos1ex01_1.log master
```

```
solve vgate=0 vstep=0.25 vfinal=3.0 name=gate
```

```
save outf=mos1ex01_1.str
```

In the above commands we are solving the drain current for Gate Voltage being given a Ramp Voltage starting from 0V and terminating at 3V incrementing in 0.25V steps.

The result is saved as a string.

```
# plot results

tonyplot mos1ex01_1.log -set mos1ex01_1_log.set
```

Now we plot the transfer characteristics at a constant drain voltage of 0.1V.

```
# extract device parameters

extract name="nvt"
(xintercept(maxslope(curve(abs(v."gate"),abs(i."drain"))))
\

- abs(ave(v."drain"))/2.0)

extract name="nbeta"
slope(maxslope(curve(abs(v."gate"),abs(i."drain"))))
\

* (1.0/abs(ave(v."drain")))

extract name="ntheta" ((max(abs(v."drain")) *
$"nbeta")/max(abs(i."drain"))) \

- (1.0 / (max(abs(v."gate")) - ($"nvt")))
```

```
quit
```

The above statements extract the Threshold Voltage of MOS Transistor, Beta and Theta.

Threshold voltage is extracted by calculating the maximum slope of Id/Vgs characteristics, extending the tangent drawn at the maximum slope point to make an intercept on Vgs axis. The X-intercept minus the half of drain voltage is Threshold Voltage.

In this extract statement, nbeta is the intercept on X-axis.

Similarly ntheta is the final value of Threshold Voltage.

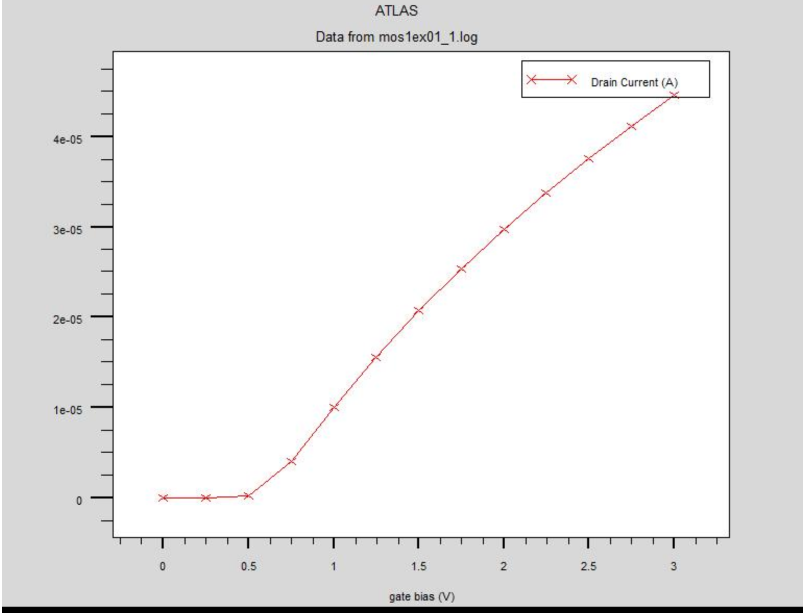


Figure 16. Plot of I D -V GS at constant V DS and extraction of V Th .

The Tony Plot of the transfer characteristics of the simulated device is given in Figure 16.The whole program has been deciphered and explained in the simplest terms. Following the same set of statements similar simulations can be carried out.

SSPD_Chapter 7_Syllabus of CMOS DIGITAL VLSI DESIGN

CMOS DIGITAL VLSI Design is being offered as Elective III in four years B.Tech Course of Electronics and Communication Engineering. This module contains the syllabus of the same.

EC1XXXX_Elective III_CMOS Digital VLSI Design[3-1-0]

Introduction to IC Technology- Introduction to IC Technology, MOS and related VLSI technology, basic MOS Transistors(enhancement and depletion mode), NMOS process, CMOS process(P-well, N-well and twin-tub), BICMOS process flow, aspects of CMOS and BiCMOS devices. 6 L

Basic Electrical Properties of MOS Circuits – MOSFET Threshold Voltage, I-V relationships for MOSFET, MOSFET Transconductance, the pass transistor, NMOS inverter, pull-up to pull-down ratio for NMOS inverter driven by NMOS inverter and pass transistor, different forms of pull-up(load resistor, depletion mode NMOS, enhancement mode pull up, CMOS pull up), CMOS Inverter, MOS transistor circuit model, Latch-up in CMOS circuits, BICMOS Inverter, Comparative aspects of CMOS and BiPolar Transistors. 7 L

MOS circuit design processes – MOS layers, stick diagrams(NMOS design style, CMOS design style), Euler Path and Design Optimization, Design rules and layout (Lambda-based design rules, contact cuts, double metal MOS process rules, CMOS – lambda based design rules), 2 micron double metal double poly CMOS rules. 5 L

Basic circuit concepts – sheet resistance, area capacitance of layers, inverter delays, driving large capacitive loads, propogation delays (cascaded pass transistor, design of long polysilicon wires), wiring capacitances (fringing fields, interlayer capacitance, peripheral capacitance) 6 L

Scaling of MOS Circuits – scaling models and scaling factors (gate area, gate capacitance, channel current density, channel resistance, gate delay, maximum operating frequency, saturation current, current density, power dissipation), limitations of scaling. 4 L

Subsystem design and layout – Switch logic (pass transistors and transmission gates), Gate logic (inverter, 2-input CMOS NAND and NOR gates), structured design of a parity generator. 4 L

Memory and aspects of system timing – System timing consideration, 1-transistor dynamic memory cell, 3-transistor dynamic RAM cell area, dissipation, volatility), RAM arrays. 4 L

Practical aspects – Optimization of NMOS and CMOS inverters, I/O pads, aspects of design tools (graphical entry layout, design verification, design rule checkers, circuit extractors, simulators). 4 L.

Text books: i. Basic VLSI Design – Pucknell and Eshraghian;

ii. VLSI Fabrication principles – Sorab Gandhi;

Reference books: i. The science and engineering of Microelectronic Fabrication – Stephen Camplell;

ii. VLSI Design – Sujata Pandey and Manoj Pandey;

iii. CMOS VLSI Design – Wolfe.

SSPD_Chapter 7_Part 1_The Art of Microchip Fabrication transforms into HiTech Computer-Aided VLSI DESIGN and Microchip Fabrication.(Untitled) SSPD_Chapter 7 is concerned with the physical level design of VLSI Circuits. SSPD_Chapter 7_Part 1 deals with the level of Automation which has been achieved in the design of VLSI circuits.

SSPD_Chapter 7_Part 1_The Art of Microchip Fabrication transforms into HiTech Computer-Aided VLSI DESIGN and Microchip Fabrication.

[This chapter is based on the write up titled ‘Technology Computer Aided Design for Si, SiGe and GaAs Integrated Circuits’ edited by G.A. Armstrong and C.K. Maiti as a part of IET Circuits, Devices and Systems Series 21]

In last 60 years, the successful scaling of the silicon devices and advances in processing technologies has enabled us to achieve dizzying heights in terms of complexity of electronics systems integrated till date. The advanced level of integration and complexity achieved is testified by 3D HDTV and 4G Mobile Communication.

Today we have successfully marched from micro era to nano era of IC Technology and reached giga scale level of integration but we are running into serious road blocks . These road blocks are due to reaching the physical limits of the devices and quantum mechanics will have to be invoked for further scaling of the devices.

- Principles of VLSI Design

[CMOS VLSI Design: A Circuits and Systems Perspective, Third Edition.by Neil H.E. Weste and David Harris.ISBN: 0-321-14901-7, Addison Wesley.

Supplementary texts : Digital Integrated Circuit Design by Ken Martin, Oxford University Press (2000). Digital Integrated Circuits, A Design Perspective, Second Edition by J. Rabaey, A. Chandrakasan and B. Nikolic, Prentice Hall (2003).]

What are the concepts and techniques of modern integrated circuit design (CMOS VLSI).

How to design integrated circuits using Commercial Computer Aided Design (CAD) Tools (CADENCE).

The Design Process:An iterative process that refines an *idea* to a manufacturable device through at least five levels of design abstraction.

Top level: The *idea* refined into a set of requirements called specifications:

What does the chip do?

How fast does it need to operate in order to be competitive?

How big will it be?

How much power will it consume?

Design Constraints:

Speed, power and area.

Abstraction:A very effective means of dealing with design complexity.

Creating a model at a higher level of abstraction involves replacing detail at

the lower level with simplifications.

Simulation:The functional behavior of the design (or a parameter such as power)

is determined by applying a set of excitation vectors to a circuit model.

Levels of abstraction:

(1) Functional (architecture)

(2) Register Transfer Level (microarchitecture, block)

(3) Logic Design

(4) Circuit Design

(5) Physical Design

Hardware Description Languages

Verilog, VHDL etc.

VHDL Example: 32 bit adder

```
Entity ALU32 is port(

A, B: in bit_vector(31 downto 0);

.....) end ALU32;

if (a=b) then

sum < = ‘0’ ;

else

sum < = (a or b);

end if;
```

This is Behavioural Simulation.

In the following four diagrams the four levels of abstraction are introduced.

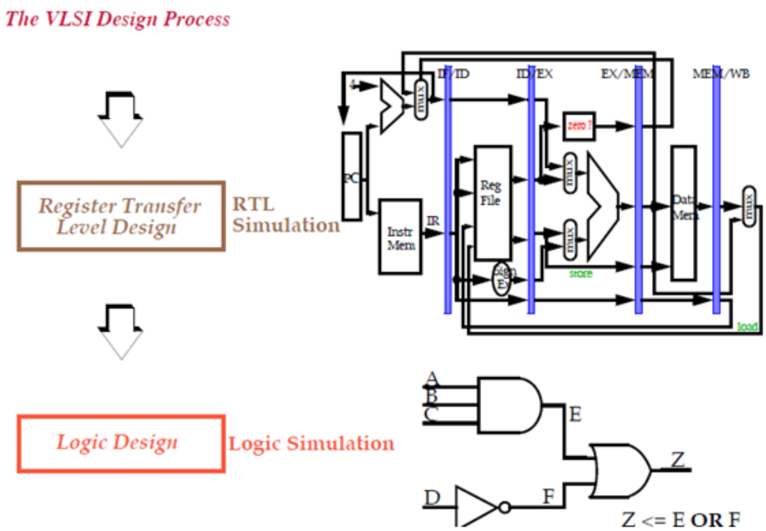


Figure 7.1. VLSI Design at RTL and Logic Design Level.

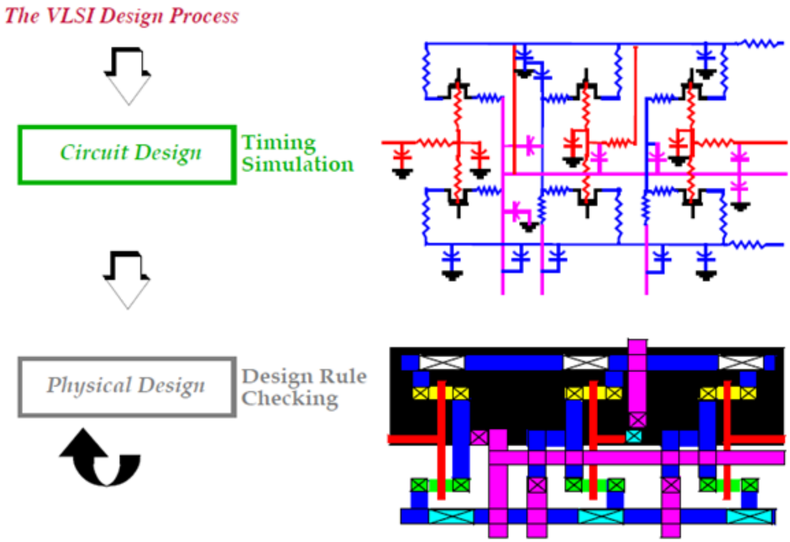


Figure 7.2. VLSI Design at Circuit Design and

Physical Design Level.

In this SSPD_ Chapter 7 we are concerned with physical design level.

The layout of a CMOS inverter is shown below. In course of the discussion of Chapter 7 we will arrive at the particular layout of CMOS as shown in Figure 7.3

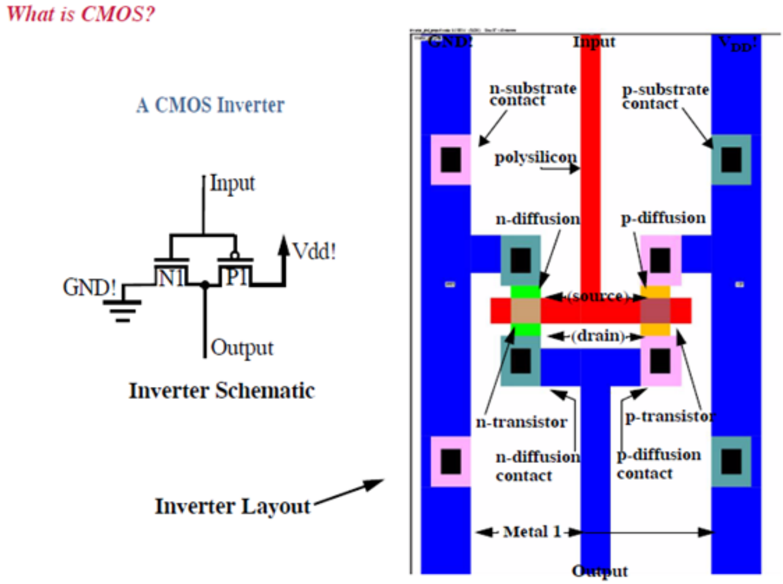


Figure Prin Figure 7.3. CMOS Layout on IC Chip. CMPE 315

7.2. Hierarchy and Abstraction

Moore's Lawstates: Integration density doubles every 18 months.

For example, Microprocessors:

The million transistor/chip barrier crossed in ‘88 with the 486. Today we have more than 100 million transistors on a single chip in Pentium IV..

Impact of this revolution on design is:

Hand crafting not possible anymore for designing a pentium IV (as was done for the 4004).

Hierarchyis used in the design of complex VLSI circuits.

A large system can be partitioned into many units. Each unit can have functional blocks,

Functional blocks are built from cells, cells are ultimately constructed from transistors.

The processor is a collection of modules each composed of cells. Re-use of cells reduces design effort.

In Figure 7.4, the hierarchical levels of VLSI Design is shown. It is top-down approach.

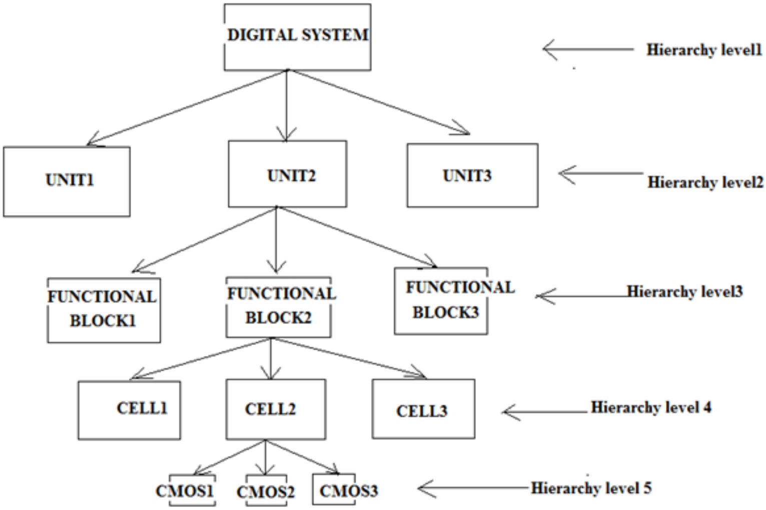


Figure 7.4. Hierarchical Level Representation of VLSI Design.

Abstractionis also used in digital designs. It is critical for dealing with the design complexity.

Entire CAD design frameworks are based on this design philosophy.

These have made it possible to achieve current design complexity.

Examples of CAD tools for digital design are:

Simulators that work at various complexity levels.

Design verification tools.

Place and Route tools. (Layout generation)

Logic synthesis tools.

Standard cellsare a popular design style that makes layout generation easy.

Layouts of basic gates such as AND, OR, NAND, NOR, and NOT as well as

arithmetic and memory modules are provided as input.

These cells are designed with similar characteristics, such as constant height,

and can be manipulated easily to generate a layout.

Place-and-Route tools can use these libraries and generate layouts using logic level description of the design.

Digital Circuit Design

If design automation solves all the problems, why be concerned with digital circuit

design?

Reality is more complex and a knowledge of digital circuit design will be important for some time to come.

Someone has to design and implement the module libraries.

Porting from technology generation to technology generation (different feature sizes) is NOT automatic.

This occurs approximately every two years!

□ Creating an adequate *model* of a cell/module requires an in-depth understanding of its internal operation.

□The library-based approach does NOT work for all situations, i.e. high performance sub-systems in designs like microprocessors.

The abstraction-based approach is only correct to a certain degree.

Performance of a module, i.e. an adder, is substantially influenced by the way it is connected in its environment (*interconnect parasitics*).

□Scaling tends to emphasize other deficiencies of the abstraction-based approach.

Global entities, such as clock signals and supply lines, are significantly

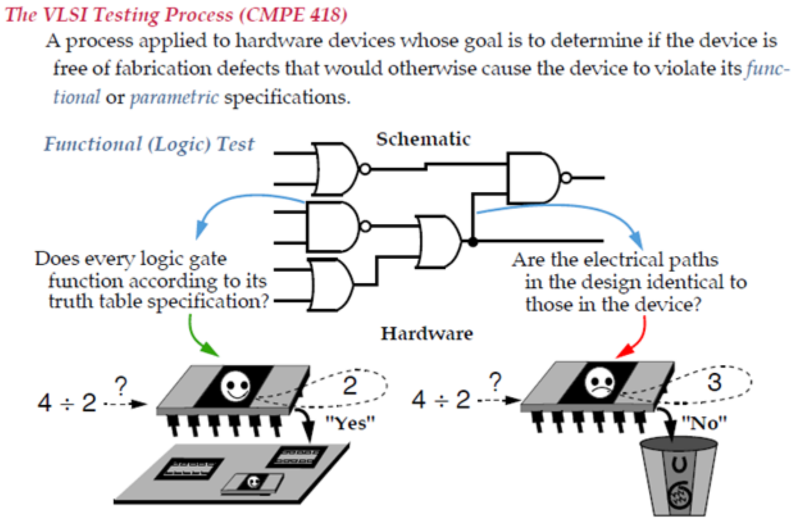
affected by scaling.

☐ New design issues emerge over time.

Power dissipation issue periodically re-emerges.

☐ Trouble shooting an erroneous design requires circuit expertise.

☐**And**: You need to know it for doing the class assignments and project.



The VLSI Testing Process (CMPE 418)

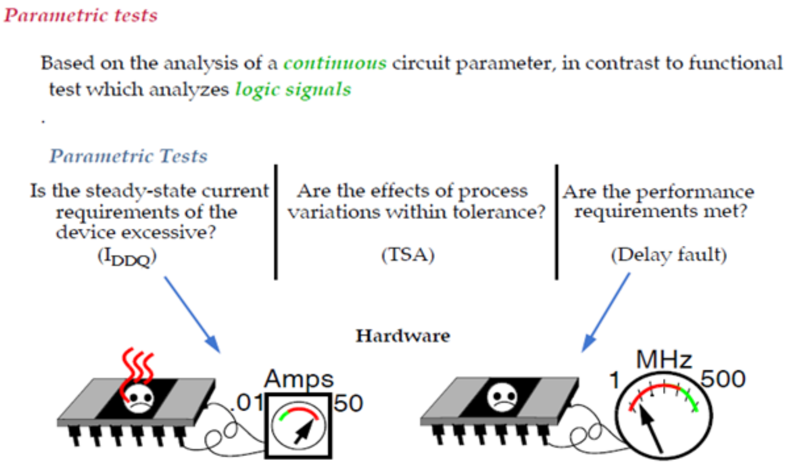
A process applied to hardware devices whose goal is to determine if the device is free of fabrication defects that would otherwise cause the device to violate its *functional* or *parametric* specifications.

Functional (Logic) Test

Does every logic gate function according to its truth table specification?

Are the electrical paths in the design identical to those in the device?

Figure 7.5. Functional testing of the VLSI System.



Parametric tests

Based on the analysis of a *continuous* circuit parameter, in contrast to functional test which analyzes *logic signals*

.

Parametric Tests

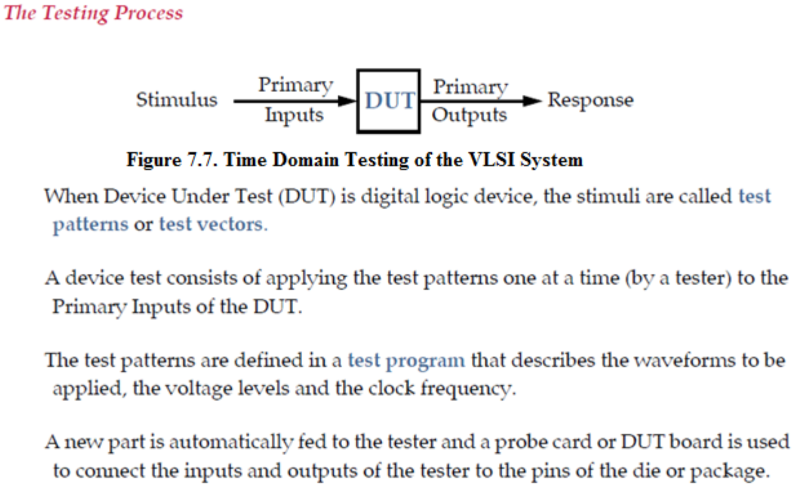
Is the steady-state current requirements of the device excessive?
(I_{DDQ})

Are the effects of process variations within tolerance?
(TSA)

Are the performance requirements met?
(Delay fault)

Hardware

Figure 7.5. Functional and Parametric Testing.



The Testing Process

Stimulus
Primary Inputs
DUT
Primary Outputs
Response

Figure 7.7. Time Domain Testing of the VLSI System

When Device Under Test (DUT) is digital logic device, the stimuli are called test patterns or test vectors.

A device test consists of applying the test patterns one at a time (by a tester) to the Primary Inputs of the DUT.

The test patterns are defined in a test program that describes the waveforms to be applied, the voltage levels and the clock frequency.

A new part is automatically fed to the tester and a probe card or DUT board is used to connect the inputs and outputs of the tester to the pins of the die or package.

7.3. History of Chip Design

[“The History of Custom MOS Design”, Steve Golson,Trilobyte Systems,Email:

sgolson@trilobyte.com

Web: http://www.trilobyte.com]

1968 – This was the beginning Custom MOS design.

- Hand-drawn schematics

- Standard cells -- new library for each product line

~ 6 kinds of gates, ~ 4-5 speed ranges

- Gate sizing by hand using CT curves (load vs delay)

- Place & route by hand on rubylith mylar sheets

- TTL breadboard for verification

- Transistor-level timing analysis using SPICE-like programs

Designer thinks about:

Gates, transistors, breadboard, and maybe library

Early 1970s -- some improvements

+ Module-level design (multiple designers working on one chip)

+ Software simulation (cycle-based)

+ On-chip bus design (trade wires for time)

+ Pitch-matched layout (datapath, PLA, ROM, RAM)

+ MOS-specific circuit techniques (dynamic logic)

Designer thinks about:

Gates, netlist simulation, floorplan, transistors

Early 1980s -- LSI Logic et al.

Hand-drawn schematics

- Hand-typed netlist

+ Automatic place and route

- Netlist simulation

+ Gate-level timing analysis

Designer thinks about:

1. Gates, netlist simulation
2. Mid 1980s -- some improvements

Schematic capture

+ Automatic netlist generation

- Automatic place and route

- Netlist simulation

- Gate-level timing analysis

Designer thinks about:

1. Gates, netlist simulation
2. Early 1990s -- Verilog and synthesis

RTL Verilog

+ RTL simulation

+ Synthesized netlist

- Netlist simulation

- Automatic (?) place and route

- Gate-level timing analysis

Designer thinks about:

1. RTL code, RTL simulation, gates, netlist simulation
2. Late 1990s -- Behavioral synthesis

Behavioral Verilog, some generated by graphical HLDA tools

+ Behavioral Verilog and HLDA simulation

- RTL Verilog, some automatically generated by HLDA tool

- RTL simulation

- Synthesized netlist

- Netlist simulation

- Automatic (???) place and route

- Gate-level timing analysis

  1. + FPGA breadboard

Late 1990s -- Behavioral synthesis

Designer thinks about:

Behavioral code, RTL code, gates, floorplan,

  1. behavioral simulation, RTL simulation, netlist simulation, breadboard

2001 and beyond the infinite…

+ System design language (C++, Superlog, SystemC, etc.)

+ System design language simulation

+ System design language synthesis

• RTL Verilog

• RTL simulation

• Floorplan

• RTL synthesized netlist & placement (physical synthesis)

• Netlist simulation

1. • Gate-level timing analysis

2001 and beyond the infinite…

Designer thinks about:

System code, RTL code, gates, floorplan, placement,

1. system simulation, RTL simulation, netlist simulation, breadboard
2. Levels of abstraction today…

[Steve Golson -- Trilobyte Systems -- Levels of Abstraction: The History of Custom MOS Design]

Levels of abstraction today…

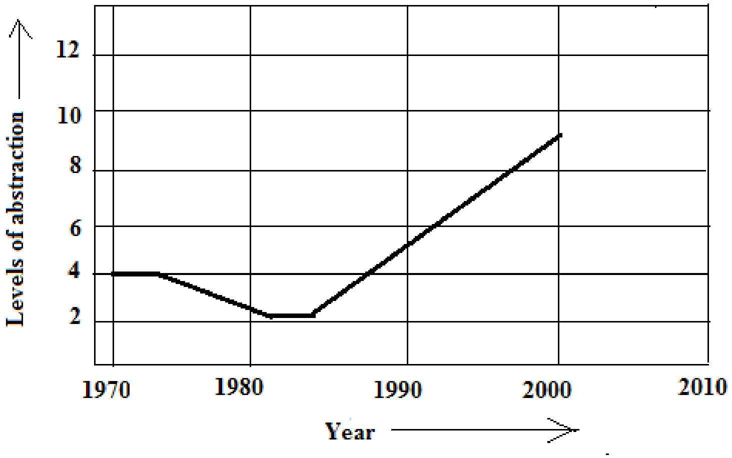


Figure 7.8. Levels of abstraction versus Year.

Levels of abstraction tomorrow??

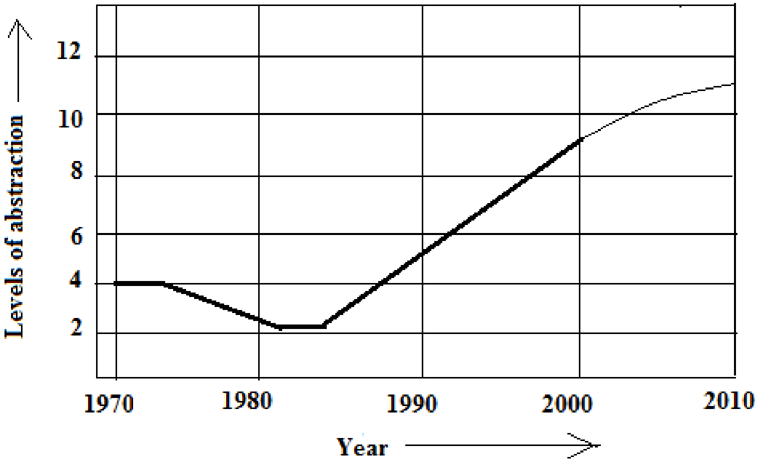


Figure 7.9. Levels of abstraction in Future.

What designers need

is to have abstraction levels *removed*,

not *added*

What designers need… is RTL signoff

Designer thinks about:

System code, RTL code,

System simulation, RTL simulation

Vendor thinks about:

everything else

Levels of abstraction: RTL signoff

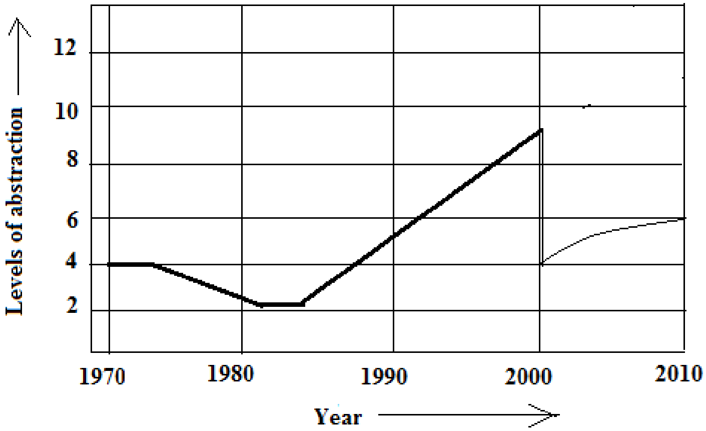


Figure 7.10. With RTL signoff, level of abstraction comes down to 4 level. As

complexity increases the levels again start adding.

What designers need…

- Physical synthesis
    - + Improve quality of synthesis
    - + Improve quality of place & route
    - + No more gate tweaks and layout modules
    - + No longer worry about gate-level timing
- Automatic verification of RTL vs netlist
    - + No longer worry about netlist simulation
- RTL analysis tools for timing, power, area
- Smarter IP lawyers

Things to consider

about new EDA tools

- How do they fit into my existing design flow?
- Revolutionary or evolutionary?
- What problems do they solve?

• What new problems might they introduce?

Higher-level abstractions place you further away from

an intuitive feel for hardware and timing: “How fast is it?”

Just because you can do something,

doesn’t mean it is worth doing

- Chip Design.

These are carried out at five levels of abstraction:

1. System level;
2. Logic level;
3. Circuit level;
4. Layout level;
5. Technological level.

Chip Design starts with product specification followed by front end design and back end design.

Front end design starts from system level and goes to the lower level using top-down approach. Up to logic level we are not concerned if we are using Bipolar Technology or MOS technology. After the logic level we come to circuit level. At circuit level, the logic functionality, timing delays, speed and power are the primary concerns. This level is technology dependent it is relatively process-independent.

At the back-end, the final design needs to be translated into the physical layout. This SSPD_chapter 7 is concerned with the physical layout.

At the front-end we have electronic computer-aided design(ECAD). These are so powerful today that the logic design can be synthesized from a high level description language such as VHDL or Verilog. The circuit netlist can be extracted from the logic functional description and the layout can be extracted from the circuit and logic-level. This part is full well explained in Collection on Digital System Design by the same author.

- History of Device Simulator.

In early 1970, numerical processor was coupled with device simulators in SUPREM and SEDAN at Stanford University. These were 1-D programs.

SEDAN could analyze devices with five metallurgical junctions. Here device analysis was carried out by solving Poisson Equation and transient current continuity equation. This gave the potential distribution and carrier concentration distribution in Silicon devices. In Table 7.1 we give the chronological development of Device Simulator.

Table 7.1. Chronological Development of Device Simulator.

Year	Simulator	Function
1970	SUPREM,SEDAN	1-D programs run to obtain potential distribution and carrier concentration.
1980	MINIMOSBAMBI,PISCES,BIPOLE,HQUPE	2-D numerical simulators for solving 2-D Poisson's Equation and current continuity equation.MINIMOS source codes were made widely available and was specially developed for MOS structures.
1980	IBM'sFEDSS/ FIELDAYAgere'sPROPHET,PADRE;	2-D simulators for solving 2-D Poisson's

		Equation and current continuity equation.
Late 1980	MINIMOS-NT	This handles planar and non-planar device structures in 2- and 3-Dimensions.
1980s	BIPOLE	Quasi-2-D device simulator. Input consists of mask dimensions, impurity profiles and carrier life-time. It solves for the terminal electrical characteristics. Extremely fast in computing time. It models Avalanche Multiplication and Avalanche Breakdown in fast transistors. It allows for parameter extraction.

Late 1980s	PISCES	General purpose 2-D simulator for Bi-pole and MOS structures. It was adopted by SILVACO
1990	ATLAS	Silvaco International Product, originally derived from PISCES, has enhanced capability of carrying out numerical, physics-based,2- and 3-D simulation of semiconductor devices.
1990s	Coupling of ECAD with TCAD	To support the process and device engineers in exploring the design space, provisions have been made to couple Electronic Design and Technology

Design.

7.6. History of Process Simulators.

Process Simulator models all the fabrication steps encountered in IC Chip fabrication such as: oxidation, implantation, diffusion, etching, deposition and lithography.

Early models depended on analytical equations. But with smaller dimensions, second order effects became significant. Hence numerical analysis became inevitable.

With shrinking dimensions, according to Moore’s law, 3-D analysis of devices has become important. Later versions of SUPREM have these capabilities.

7.7. Evolution of TCAD.

TCAD modeling based on computer simulation spans the interrelated disciplines of circuit design, device engineering, process development and integration into manufacturing. TCAD is routinely used for process and device development.

TCAD is used for rapid prototyping as well to study new device structures and discover newer terminal characteristics.

SUPREM for process simulation and PISCES for device simulation have been commercialized by Technology Modeling Associates(TMA) by the name TSUPREM4 and MEDICI respectively. Silvaco bought the license and offered the same by the name ATHENA and ATLAS. Integrated Systems Engineering(ISE) offered its own version of process and device simulator by the name DIOS and DESSIS. ISE has since merged with Synopsis.

SSPD_Chapter 7_Part 2_BiCMOS Technology

SSPD_Chapter 7_Part 2 gives the necessity of introducing BiCMOS in CMOS circuitary and the technology of fabrication of NPN BJT n CMOS environment.

SSPD_Chapter 7_Part 2_BiCMOS Technology

7.2 In SSPD_Chapter 6_Part 6 we have extensively discussed CMOS Process Flow and CMOS fabrication technology. Here we deal with BiCMOS Technology. In Figure 7.2.1 we give the speed-power performance of available technologies.

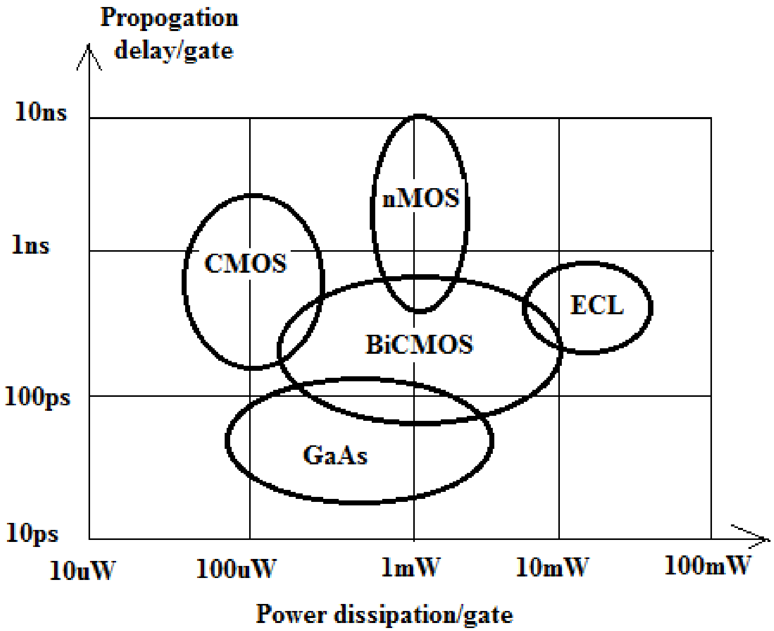


Figure 7. 2.1 Speed/power performance of available technologies.

Here we will do a rapid review of the Logic Families available in the market. Figure 7.2.2 gives the broad classification of the Logic Families. From the graph in Figure 7.2.1 it is evident that GaAs is the fastest Logic Family and CMOS is the slowest but consumes nW power and has a very high packing density hence CMOS is the technology of choice for microprocessor and microcontroller.

CMOS has one serious drawback and that is its limited driving capability. To overcome this problem one could use CMOS with super-buffers using MOS transistors alone. But still it does not compare well with BiCMOS in output drive circuits.. As seen in Figure 7.2.1, BiCMOS is faster than CMOS though dissipating more power.

CMOS and Bipolar Technology have their own advantages as shown in Table 7.2.1. BiCMOS combines the virtues of both the technology.

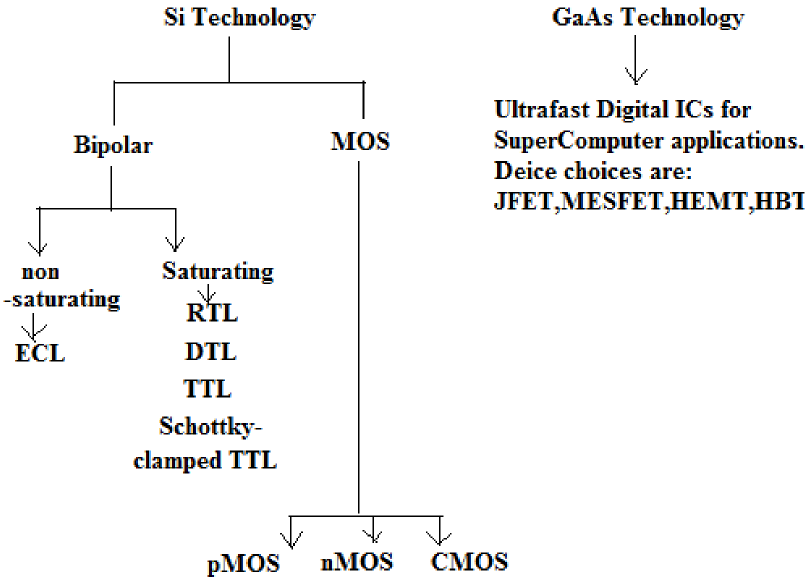


Figure 7.2. 2 The various LOGIC FAMILIES available in the market.

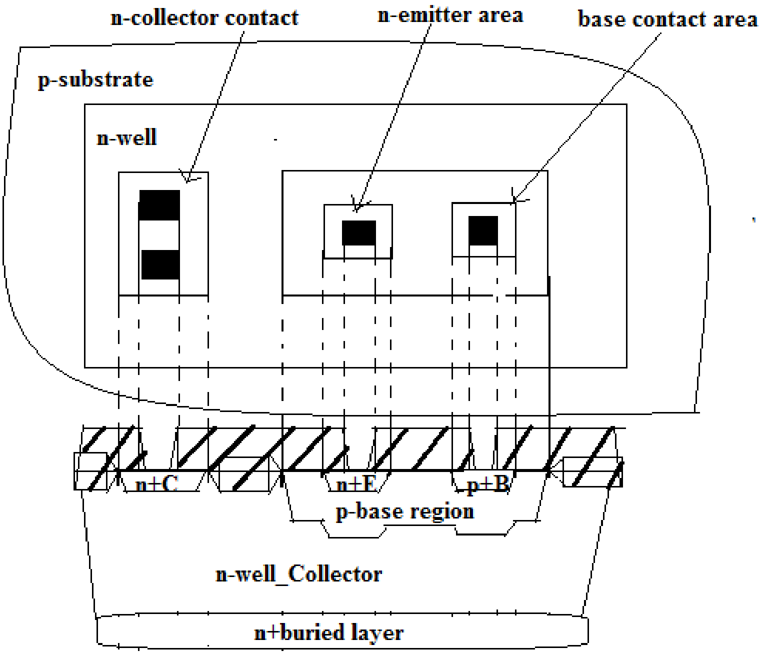
Table 7.2.1. Comparison between CMOS and Bipolar Technologies.

CMOS Technology	BJT Technology
Nano-watt logic. It has nano-watt standby power dissipation	Milli-watt standby power
High Zin hence low drive current	Low Zin hence high drive current
Scalable Threshold	Low Voltage swing

Voltage	
Noise Margin is large	Noise Margin is small
High packing density	Low packing density because of isolation diffusion
Low output drive current	High output drive current.
Low transconductance, gm	High gm
Low transit frequency	High transit frequency at low currents
Bidirectional capability- source and drain are interchangeable	Unidirectional from input to output
Near Ideal Switch	

7.2.1. Process Technology of BiCMOS.

The plan view and the cross-sectional view of a BiCMOS npn transistor is shown in Figure 7.2.3.



Black rectangles are Collector, Emitter and Base contact pads.
Hatchings are SiO2

Figure 7.2.3.BiCMOS npn transistor in n-well

CMOS process flow can easily be extended to include bipolar NPN transistor. In the twin-tub process we will use N-tub for realizing NPN BJT. To reduce the collector series resistance, n+ buried layer is provided just below the n-tub. In addition p-base region is implanted in n-well region. Next n+ implantation is done for collector contact and for emitter region. Finally p+ is implanted as shown to make contact to the base region. Next contact pads are added to the collector contact area, emitter area and base contact area as black rectangles.

Including NPN BJT causes a loss of packing density and higher cost per gate.

By using this mix of technology we can realize complex systems in an optimal manner. In the mix, we have the option of CMOS, BiCMOS and ECL.

CMOS will be confined to logic circuitary.

BiCMOS for input/output and driver circuits.

ECL for critical high speed parts.

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOS and MOS circuits to be continued1

SSPD_Chapter 7_Part 3 gives the basic elecrical properties of MOS devices.

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOS and MOS circuits.

7.3 The Output Characteristics of (Depletion)n-channel MOSFET.

Field Effect Transistors cover JFET and MOSFET. The underlying principle of FET is that there is a conducting channel which obeys Ohm’s Law:

$$I_{DS} = \frac{V_{DS}}{R_{channel}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 7.3.1$$

Where

$$R_{channel} = \frac{\rho \times L}{A}$$

L = lemgth of the channel and A = cross-sectional area as shown in Figure 7.3.1.

And ρ = resistivity = 1/σ = 1/conductivity = 1/ $(q\mu_n n)$;

Here q = charge of an electron, μ_n = mobility of an electron in the channel andn = conducting electron density in the n-channel.

Generally the n-channel is like a 2-D sheet hence its mobility is much lower than that in 3-D bulk. Typically μ_n = 1450 cm2/(V-sec) in Bulk Semiconductor and 650 cm2/(V-sec) in the channel in FET.

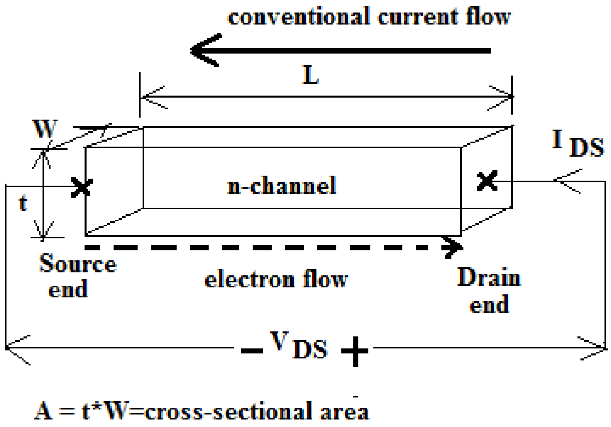
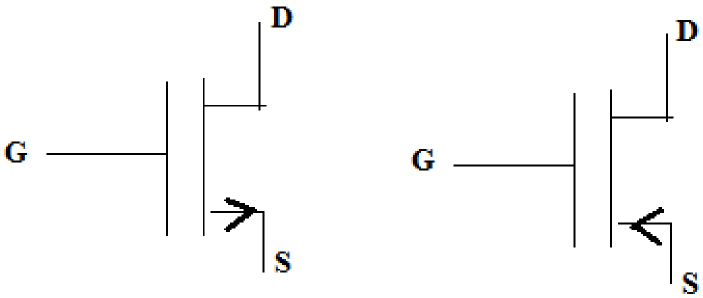


Figure 7.3.1. Current flow through n-channel of length L and cross-sectional area A = t*W where t is the thickness and W is the width of the channel.

.

As can be seen in the Figure 7.3.1 in n-channel FET electrons are being sourced from the source end and being drained out from the drain end. By corollary in p-channel FET holes will be sourced and holes will be drained out. This is the reason why the symbols are shown with the arrow of the conventional current flow. In nMOSFET conventional current is shown to come out of the Source and in pMOSFET conventional current is shown to enter into the Source as shown in Figure 7.3.2.



n-MOSFET (depletion) p-MOSFET(depletion)

Figure 7.3.2 Symbols of two kinds of MOSFETs

Equation 7.3.1 is a linear equation with I-V characterisyics as shown in Figure 7.3.3.

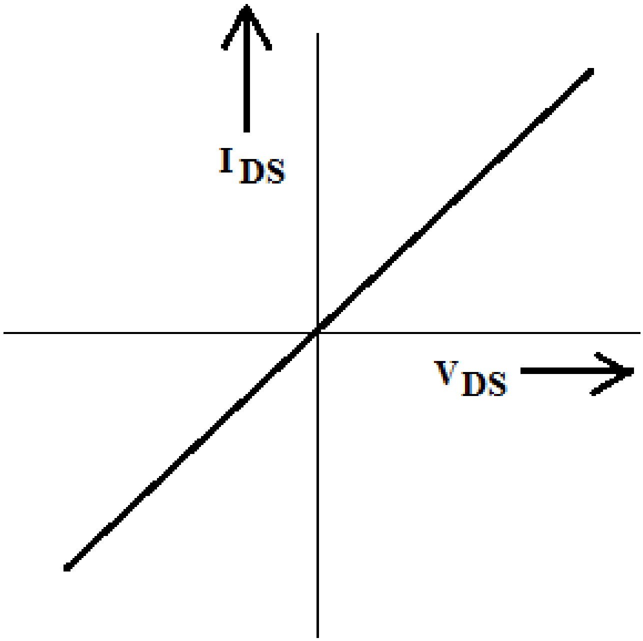


Figure 7.3.3. I-V characteristics of n-channel.

As seen in the Figure 7.3.3, I-V is linear for the whole range of voltage applied and is anti-symmetrical about the y-axis. Hence n-channel is a bilateral device where for positive voltage positive current flows and for negative voltage negative current flows with equal ease.

7.3.1. FET as a 3-terminal Device in Ohmic Region/ Triode Region

Now by applying GATE voltage V_{GS} transverse to the channel as shown in Figure 7.3.4 the channel thickness can be controlled and thereby the I-V slope can be controlled as shown in Figure 7.3.5.

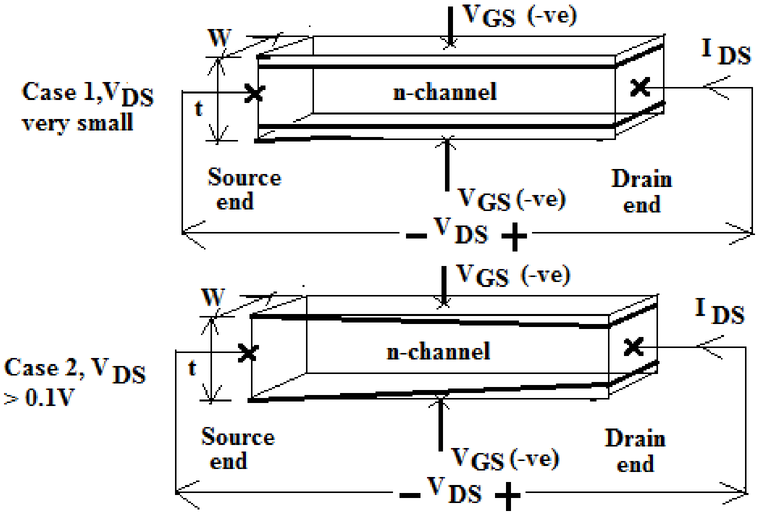


Figure 7.3.4. Modulation of the the cross-sectional area by applying negative gate voltage transverse to the channel. When V_{GS} = 0V then the channel is a parallaelopied with uniform cross-sectinal area along the length of the channel. When V_{GS} = -1V and V_{DS} < 0.1V then also channel is parallaelopied but with narrower cross-sectional area hence it offers higher resistance. When V_{GS} = -1V and V_{DS} is larger than 0.1V then the channel becomes trapezoidal and it doesnot offer a linear resistance.

In this OHMIC REGION of operation, we have two cases. First case is when V_{DS} is kept less than 0.1V. In this case channel remains a parrallaelopied with a reduced thickness ‘t’. In this case if we continue to apply more and more negative voltage, the channel will be eventually be completed depleted and we say that the channel has pinched off. The negative gate voltage at which this occurs is known as Pinch-Off voltage (V_P).

There is a second case where V_{DS} is greater than 0.1V and it continues to be increased. In this case the difference voltage between Gate and Drain > the difference between Gate and Source. Hence n-channel is trapezoidal and it does not give a linear I-V characteristics. This is also OHMIC REGION but non-linear. Linear and non-linear I-V curve is clearly indicated in Figure 7.3.5.

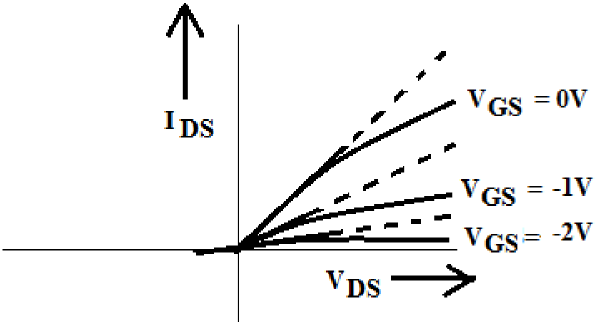


Figure 7.3.5 I-V characteristics of n-channel MOSFET (Depletion) with the application of third terminal negative Gate Voltage transverse to the channel.

As seen in Figure 7.3.4 and Figure 7.3.5 we have achieved Voltage –Controlled Resistance (VCR) but this VCR is not quite linear. At $V_{DS} < 0.1V$, VCR is perfectly linear but at large V_{DS} voltages the channel resistance deviates from the linear behavior.

In this kind of device at zero gate voltage, channel is present. Hence it is known as Normally-On . As the negative gate voltage is made more negative, the channel depletes until it finally completely pinches off . Hence it is known as .Depletion Type Device.

As the gate voltage is made positive, channel becomes more conductive as it has a copious supply of electrons. Theefore we say that n-channel has enhanced.

All the above cases are Ohmic or Triode Region of operation of (D) n-channel MOS.

7.3.2. Saturation or Pentode Region of operation of (D) n-channel MOS.

For a given V_{GS} , as V_{DS} is increased a point comes when $(V_{DS}-V_{GS}) = V_P$. At this point the channel gets pinched off on the Drain-side and n-channel becomes conical as shown in Figure 7.3.6. At this point let $V_{DS} = V_{DS}^*$. Now the current I_{DS} saturates.

Now if we increase $V_{DS} = V_{DS}^* + \Delta V$, the apex of the conical channel becomes more pinched off and the incremental voltage ΔV drops across the pinched off region between the apex of the cone and the drain. The voltage drop across the conical channel remains fixed at V_{DS}^* and the resistance of the conical channel is also constant at 1/3 of the full value of the parallaelopied channel resistance. Thus the drain current saturates at :

$$I_{DS(sat)} = \frac{V_{DS}^*}{\left(\frac{R_0}{3}\right)}$$

Where R0 = the resistance of the parallaelopied channel.

The saturation region of operation or Pentode Region of operation is shown in Figure 7.3.7.

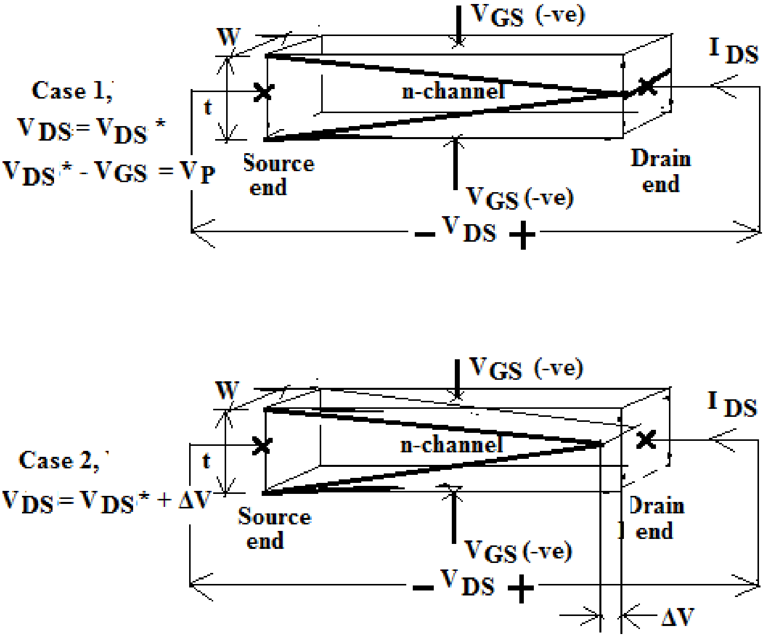


Figure 7.3.6. (D) n-channel MOS operating in Saturation Region or Pentode Region.

All these arguments hold good for Enhancement Type MOSFETS also except that (E)nMOS is normally-off device and (D)nMOS is normally – on device. In Enhancement type NMOS, positive voltage has to be applied to the gate and only when Gate Voltage has exceeded the Threshold

Voltage(V_{Th}) that the n-channel is induced through inversion. In contrast in Depletion type NMOS, n-channel is already present there. Gate Voltage can be negative in which case the n-channel will get depleted and ultimately it will be pinched-off. Gate Voltage can be positive . In this case channel will beome more conductive.

(D)nMOS gives a greater flexibility of design hence it more preferred device.

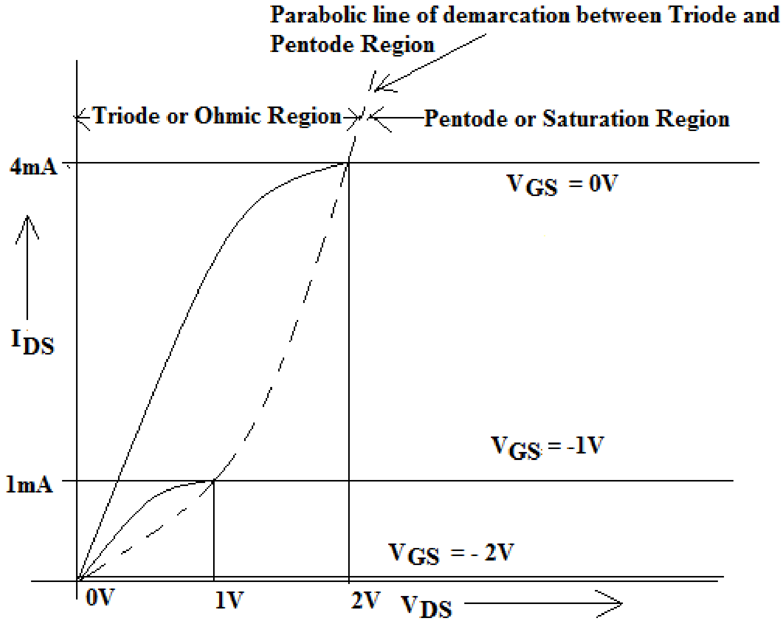


Figure 7.3.7. The output family of curves of (D)n-channel MOS demarcated between the Triode and Pentode Regions of Operations.

The Transfer Characteristics of (E)NMOS and (D)NMOS are shown in Figure 7.3.8. As can be seen from the Graph in Figure 7.3.8, (E)NMOS originates at V_{Th} whereas (D)NMOS originates at pinch-off voltage V_P . Both these transfer characteristics have been plotted in the Pentode region of the Output Characteristics. The saturation current of (D)NMOS at $V_{GS} = 0V$ is referred to as I_{DSS} as it is done in JFET.

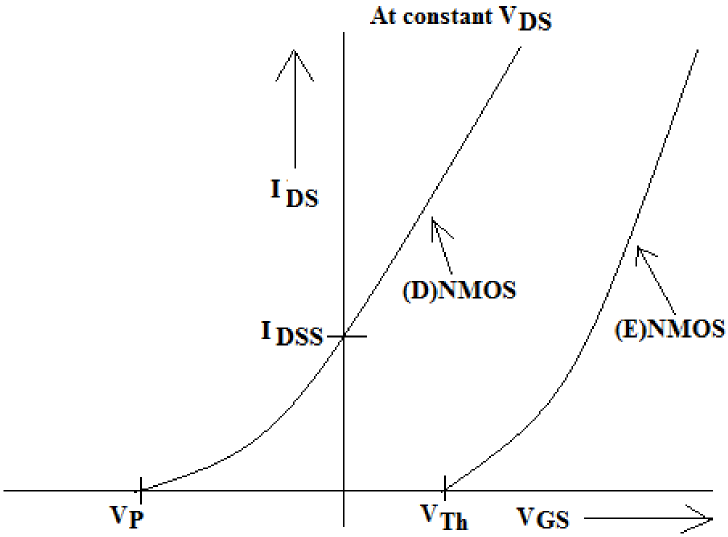
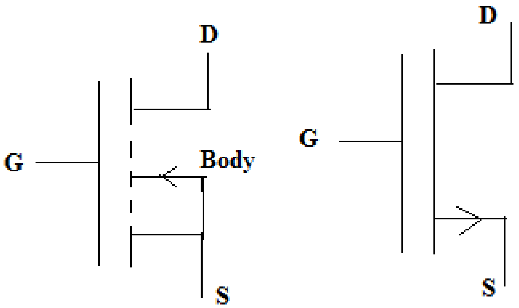


Figure 7.3.8. The Transfer Characteristics of (D)NMOS and (E)NMOS.



(E)NMOS (D)NMOS

Figure 7.3.9. Symbols of (E)NMOS and (D)NMOS. Normally-off device is symbolized by broken line depicting broken channel. Normally-on device is depicted by continuous line.
Body and Source are shorted together to minimize the body-effect.
If the substrate (i.e. the Body) is not kept at zero voltage the threshold voltage increases through body-effect. To minimize this effect substrate and source are shorted together.

7.3.3.Theoretical Formulation of Output family of Curves of NMOS.

Just as in BJT,

$$I_C = \frac{Q_B}{\tau_t}$$

Where Q_B = charge stored in Base and τ_t = transit time in the Base.

Similarly

$$I_{DS} = \frac{Q_C}{\tau} = \frac{\text{charge induced in the channel}}{\text{electron transit time in the Channel}}$$

7.3.3

Transit time has the following formulation:

$$\tau_{sd} = \frac{L}{v_{drift}} = \frac{Length\ of\ the\ Channel}{drift\ velocity\ of\ electron\ in\ the\ channel}$$

7.3.4

Drift Velocity of electron in the channel is as follows:

$$v_{drift} = \mu_n E_{ds}$$

7.3.5

Where μ_n = 2D mobility of electron in the channel because the channel behaves like a sheet and not like a Bulk.

E_{ds} = Electric field from the Drain to Source

$$E_{ds} = \frac{V_{ds}}{L}$$

7.3.6

Substituting Eq.7.3.6 in Eq 7.3.5 which in turn is substituted in Eq.7.3.4. we get:

$$\tau_{sd} = \frac{L}{v_{drift}} = \frac{L^2}{\mu_n V_{ds}} \qquad 7.3.7$$

This is analgous to the expression we get for the transit time of electron in NPN BJT:

$$\tau_t = \frac{W^2}{2D_n} = \frac{W^2}{2\mu_n V_{Th}}$$

7.3.8

Here W = effective Base width, μ_n = Bulk mobility of electron and V_{Th} = thermal voltage at room temperature = kT/q;

7.3.3.1. Output Curve in Triode Region(non-saturated region).

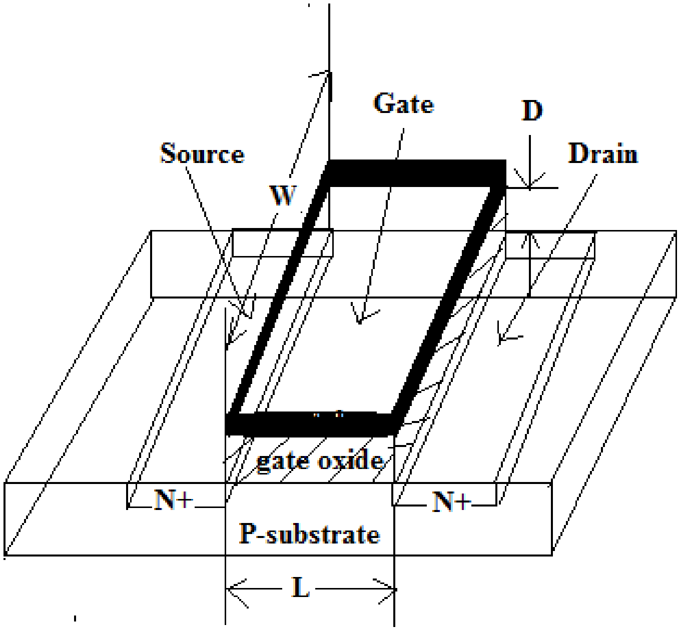


Figure 7.3.10. 3D view of NMOS transistor structure.

Figure 7.3.10 gives the 3-D view of NMOS transistor structure and it indicates the symbol used for the width, length of the channel anf for gate oxide thickness. According to the Figure:

L = length of the channel,

W = the width of the channel and D= thickness of the gate oxide.

We have to do the quantitative formulation of the n-channel charges induced in P-body after the Gate Voltage exceeds the threshold in case of (E)NMOS.

From now onward we will deal with (E)NMOS.

In (E)NMOS, positive gate voltage repels the majority carriers, namely holes, and creates depletion layer. In the process there is band-bending at the interface.

When band-bending $= 2q\psi_B$ then inversion layer is induced.

The charges on the two plates of the gate capacitance and the gate voltage is shown in Figure 7.3.11.

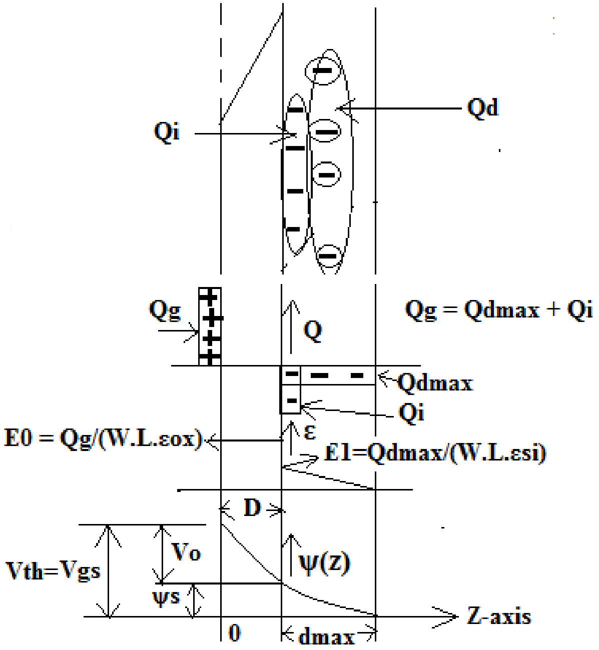


Fig.7.3.11 The charge profile, Electric Field Distribution and the Potential Drop from the Gate to bulk along the longtitudinal axis-z is shown.

Q_g (the charge on the gate plate) = Q_{dmax} (depletion layer charge at its maximum width) + Q_i(inversion layer charge).

Q_i = inversion layer charge is the channel charge which we have referred to as the charge induced in the channel = Q_C

Q_C = D_g(electric flux density terminating at the induced channel charge) $\times$WL

Total electric flux originating = Q_g .

A part is terminating on (-ve)Q_{dmax} and the remaining is terminating on (-ve)Q_C. The part which is terminating on (-ve)Q_C is accounted by Eq.7.3.9.

But $D_g = E_g \times \varepsilon_{ins} \times \varepsilon_0$

Where E_g = average gate electric field with respect to the channel,

ε_{ins} = relative permittivity of the gate oxide = 4 for Silicon Dioxide ,

ε_0 = free space permittivity or absolute permittivity = 8.854×10^{-14} F/cm.

To consider the average electric field (E_g), we must consider the average electric field of the channel.

The Drain end of the channel is at V_{ds} and the source end is at 0V. Therefore the average voltage of the channel is ($V_{ds}/2$).

Also effective gate voltage responsible for induced channel charges or responsible for inversion layer is ($V_{gs} - V_t$).

Hence average gate electric field with respect to the channel is:

$$E_g = \frac{(V_{gs} - V_t) - \frac{V_{ds}}{2}}{D}$$

7.3.10

Where D = oxide thickness.

Therefore:

$$Q_c = \left(\frac{WL \in_{ins} \in_0}{D}\right)\left[\frac{(V_{gs} - V_t) - \frac{V_{ds}}{2}}{D}\right] \qquad 7.3.11$$

Errata in Equation 7.3.11
In Equation 7.3.11., Charge induced is equal to the Electric Flux Density into total cross sectional area. Total cross-sectional area = W*L and Electric Flux Density = permittivity*Electric Field.Hence there is an extra D(thickness of Oxide Layer) in the

denominator. There is only one D in the denominator of Equation 7.3.11

But

$$I_{DS} = \frac{Q_C}{\tau}$$

And

$$\tau_{sd} = \frac{L^2}{\mu_n V_{ds}}$$

Substituting Eq.7.3.11 and Eq.7.3.8 in Eq.7.3.3 we get:

$$I_{ds} = \left(\frac{W}{L} \times \frac{\epsilon_{ins}\epsilon_0\ \mu_n}{D}\right)\left[(V_{gs} - V_t) - \frac{V_{ds}}{2}\right]V_{ds} \qquad 7.3.12$$

Gate Capacitance per unit area = C_{ox} and total gate Capacitance is C_g therefore

$$C_g = C_{ox} \times WL = \frac{\epsilon_{ox}\epsilon_0}{D} \times WL \qquad 7.3.13$$

Equation 7.3.12 is also written as:

$$I_{ds} = \left(\frac{W}{L} \times K\right)\left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right] \qquad 7.3.14$$

This equation is applicable in non-saturated region where

$$V_{ds} < V_{gs} - V_t$$

This non-saturated region is called ohmic region or triode region.

7.3.3.2. Output curves in Saturated Region (Pentode Region)

NMOS enters the saturated region when the trapezoidal channel pinches off from the drain side. This happens when

$$V_{ds} = V_{gs} - V_t = V_{ds}^*$$

7.3.15

Equation 7.3.14 reduces to:

$$I_{ds} = \left(\frac{W}{L} \times K\right)\frac{(V_{gs} - V_t)^2}{2}$$

7.3.16

When V_{ds} exceeds V_{ds*}, still the voltage drop across the conical channel from the source to the apex of the cone remains V_{ds*} and the excess voltage (V_{ds} -

V_{ds^*}) drops across the depleted region from the apex to the drain. By ballistic action electrons jump across the depleted region and maintain the current flow.

Equation 7.3.16 can be re-written as:

$$I_{ds} = \left(\frac{W}{L} \times K\right)\frac{(V_{ds}^{*})^2}{2}$$

7.3.17

As is evident from Eq. 7.3.17, the line of demarcation between Triode and Pentode is a parabola which was pointed out in the the family of output curves in Figure 7.3.7.

These equations describing the output family of curves are valid for (E)NMOS as well as for (D)NMOS except that in (E)NMOS threshold is a positive voltage typically 1V and in (D)NMOS it is a negative voltage typically -1V. Equation 7.3.14 and Equation 7.3.16 show that the electrical characteristics are critically dependent on the geometrical dimensions of the MOS devices.

SSPD_Chapter 7_Part3_Electrical properties of MOS continued2

SSPD_Chapter 7_Part 3 is the continuation of the Electrical Properties of MOS. In this module theoretical formulation of threshold voltage, transconductance, unity gain BW, output incremental conductance and figure of merit of NMOS is derived.

SSPD_Chapter 7_Part3_Electrical properties of MOS continued2

7.3.4. Theoretical Formulation of Threshold Voltage in (E)NMOS.

In SSPD_Chapter 6_Part 4_concluded_MOS and its Physics we have dealt with the theoretical formulation of the Threshold Voltage. The expression comes as follows:

$$V_{Th} = \frac{\sqrt{2qN_A\epsilon_{Si}(2\psi_B+V_{SB})}-Q_{SS}}{C_{ox}} + 2\psi_B + V_{FB}$$

7.3.18

Where The Flat Band Voltage is = $V_{FB} = \phi m - \phi s = \phi m - [\chi + E_g/(2q) + \psi_B]$

χ = electron affinity in semi-conductor, E_g = energy band-gap of Silicon and ψ_B = potemtial diiference between Fermi-level(Ef) and intrinsic level(Ei) in Si-Bulk.

Therefore $qV_{FB} = q\phi m - q\phi s = q\phi m - [q\chi + E_g/(2) + q\psi_B]$;

If source to bulk voltage (V_{SB}) is not zero then this also has to be accounted. This is called Body Effect.

$Q_{SS} = Q_{it}$ (interface trapped charges) + Q_{ot}(oxide trapped charges) + Q_m(mobile charges in the oxide) + Q_f (fixed oxide charges).

$$\sqrt{2qN_A\epsilon_{Si}(2\psi_B + V_{SB})} = Q_{dmax} \\ = \mathit{charge\ below\ the\ oxide\ interface\ in\ the\ depletion\ region;}$$

Generally substrate and source are connected together and threshold voltage is at the minimum.

But in certain applications, substrate is kept at reverse bias with respect to source. NMOS is built in P tub and PMOS is built in N tub.

Hence in NMOS, substrate is at negative bias with respect to source and in PMOS, substrate is at positive bias with respect to source.

Substrate bias gives rise to increase in threshold voltage. Hence Body-effect is represented by the following empirical expression:

$$V_{TN} = V_{TO} + \gamma(\sqrt{V_{SB} + 2\psi_B} - \sqrt{2\psi_B})$$

7.3.17

V_{TO} = Threshold voltage with zero substrate bias;

And γ = body effect parameter ($\sqrt{V}$);

$2\psi_B$ = surface potential parameter.

Lower is the substrate doping, lower will be body effect.

The Typical values of Body effect is given in Table 7.3.1.

Table 7.3.1. Body effect in (E)NMOS and (D)NMOS

Device	Substrate bias V_{SB}	Threshold V_{Th}	
(E)NMOS	0V	$0.2V_{DD}$	1V for $V_{DD} = +5V$
	5V	$0.3V_{DD}$	1.5V for $V_{DD} = +5V$
(D)NMOS	0V	$-0.7V_{DD}$	-3.5V for $V_{DD} = +5V$
	5V	$-0.6V_{DD}$	-3V for $V_{DD} = +5V$

7.3.5. Theoretical formulation of Transconductance (g m)

Transconductance is the partial derivative of output current (I_{ds}) with respect to input voltage (V_{gs}) with output voltage constant i.e. V_{DS} is held constant.

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} \Big| \; V_{DS}$$

7.3.18

From Eq.7.3.3,

$$I_{DS} = \frac{Q_C}{\tau}$$

And Transit Time is given by equation 7.3.7,

$$\tau_{sd} = \frac{L^2}{\mu_n V_{ds}}$$

Using these two equations incremental change in output current:

$$\partial I_{ds} = \frac{\partial Q_c V_{ds} \mu_n}{L^2}$$

7.3.19

But channel charge induced depends on gate capacitance and gate voltage:

$$\partial Q_C = C_g \partial V_{gs} \qquad 7.3.20$$

Therefore

$$\partial I_{ds} = \frac{C_g \partial V_{gs} V_{ds} \mu_n}{L^2}$$

Hence

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = \frac{C_g V_{ds} \mu_n}{L^2}$$

7.3.21

In saturation from Eq.7.3.15,

$$V_{ds} = V_{gs} - V_t = V_{ds}^*$$

.

$$V_{ds}^*$$

is the actual drop across the conical channel no matter what V_{ds} is.

Hence in saturatin region:

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = \frac{C_g \mu_n}{L^2}$$

(

$$V_{gs} - V_t)$$

7.3.22

From Eq.7.3.13 ,

$$C_g = C_{ox} \times WL = \frac{\epsilon_{ox} \epsilon_0}{D} \times WL$$

Substituting this in Eq.7.3.22:

$$g_m = \frac{\partial I_{ds}}{\partial V_{gs}} = \frac{C_g \mu_n}{L^2}$$

(

$$V_{gs} - V_t)$$

$=$

$$\frac{\mu_n}{L^2}$$

(

$$V_{gs} - V_t) \times \frac{W}{L} \times \frac{\epsilon_{ox}\epsilon_0}{D}$$

Or

$$g_m =$$

$$\frac{\mu_n \epsilon_{ox}\epsilon_0}{D} \times \frac{W}{L}(V_{gs} - V_t)$$

Or

$$g_m =$$

$$\beta(V_{gs} - V_t)$$

7.3.23

Where

$$\beta = \frac{\mu_n \epsilon_{ox}\epsilon_0}{D} \times \frac{W}{L} = K \times \frac{W}{L}$$

From Eq.7.3.23 it is evident that transconductance can be improved by reducing Channel Length.

Transconductance can be improved by increasing Channel Width. But both these methods have their drawbacks.

7.3.6. Theoretical Formulation of unity current gain band-width.

In BJT unity current gain band-width is defined as transit frequency ω_T . It is derived by determining the Current Gain Band Width Product.

Short Circuit Current Gain is derived as:

$$\beta = \frac{\beta_0}{1 + j\dfrac{\omega}{\omega_\beta}}$$

Where

$$\omega_\beta = \frac{1}{r_\pi(C_\pi + C_\mu)}$$

Therefore

$$\beta \times \omega_\beta = \frac{g_m}{C_\pi + C_\mu} = \omega_T$$

7.3.24

In exactly the same manner short circuit current gain of NMOS is derived and set to unity.

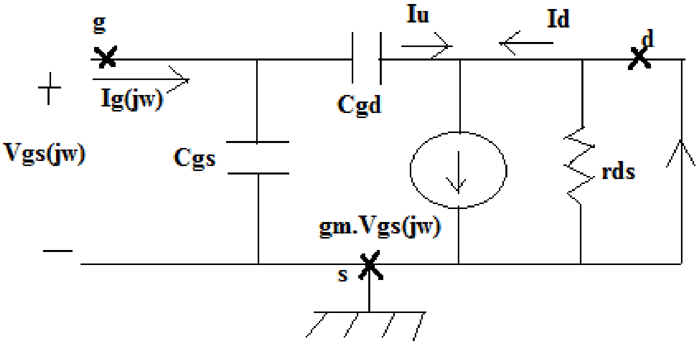


Figure 7.3.1. Small signal model of NMOS and small signal analysis for determining Unity Gain Band-Width of NMOS (w_u)

$$\frac{I_d}{I_g} = \frac{g_m\left(1 - \frac{j\omega C_{gd}}{g_m}\right)}{j\omega(C_{gs} + C_{gd})}$$

$$\frac{I_d}{I_g} = \frac{\omega_u\left(1 - \frac{j\omega}{\omega_0}\right)}{j\omega}$$

7.3.25

Where

$$\omega_0 = \frac{g_m}{C_{gd}} = this\ is\ a\ zero\ in\ Right\ Half\ Plane\ of\ s-plane.$$

$$\omega_u = \frac{g_m}{C_{gs}+C_{gd}} = Unity\ Current\ Gain\ BW.$$

In Eq.7.3.25, $\omega_0 = 5\ \omega_u$ therefore, in working range of frequencies, Equation 7.3.25 can be simplified to:

$$\frac{I_d}{I_g} = \frac{\omega_u}{j\omega}$$

7.3.26.

If Equation 7.3.26 is equated to Unity then its corresponding Unity Gain Frequency is:

$$\omega_u = \frac{g_m}{C_{gs}+C_{gd}} = Unity\ Current\ Gain\ BW.$$

7.3.27

From Eq.7.3.27 it is evident that increase in transconductance gives a higher frequency range of opeartaion.

7.3.7. Theoretical Formulation of Output

Conductance of (E)NMOS.

Theoretical expression of drain current in saturation region is given by Equation 7.3.16:

$$I_{ds} = \left(\frac{W}{L} \times K\right) \frac{(V_{gs} - V_t)^2}{2}$$

As is evident from the above Equation, I_{ds} has no dependence on V_{ds} and I_{ds}-V_{ds} family of curves are perfectly horizontal and parallel to one another. Horizontal I-V curve implies infinite output impedance of the active device. But in practice it is not so. Real MOS devices have slopes in their family of I-V curves and this slope becomes pronounced as we scale the devices for the different generation of Technology. This is known as Channel Length Modulation effect.

In BJT we have Base Width Modulation also known as Early Effect. This Early Effect is the cause of the slope in output family of curves of CB BJT and CE BJT. Due to these slopes we have h_{ob} and h_{oe} parameters in the two circuit configurations. Since Early Effect is more pronounced in CE BJT hence $h_{oe} = 1/40\text{k}\Omega$ is greater than $h_{ob} = 1/2\text{M}\Omega$. Channel Length Modulation is analogous to Early Effect and its degradation of family of output curves of NMOS is clealrly brought out in Figure 7.3.2. The physics of this degradation is that we have assumed that after saturation, I_{ds}(sat) becomes constant at:

$$I_{DS(sat)} = \frac{V_{DS}^{*}}{\left(\frac{R_0}{3}\right)}$$

7.3.2

In Equation 7.3.2 it is assumed that with increase in V_{DS}, the voltage drop across the conical channel is constant at V_{DS}^{*} and excess voltage drops across the pinched off region. The second assumption is that the resistance of the conical section is constant at :

$R_0/3$ where R_0 is the resistance of the parallelopied channel.

This assumption does not remain valid with the scaling of devices. As the the device is scaled, the variation in V_{ds} leads to significant change in the resistance of the conical channel because as pinched off region increases the axial length of the conical channel reduces and hence resistance offered decreases and $I_{DS(sat)}$ increases with the increase in V_{DS}.

dotted lines are idealized family of output curves.

Continuous lines are the family of output curves of real device.

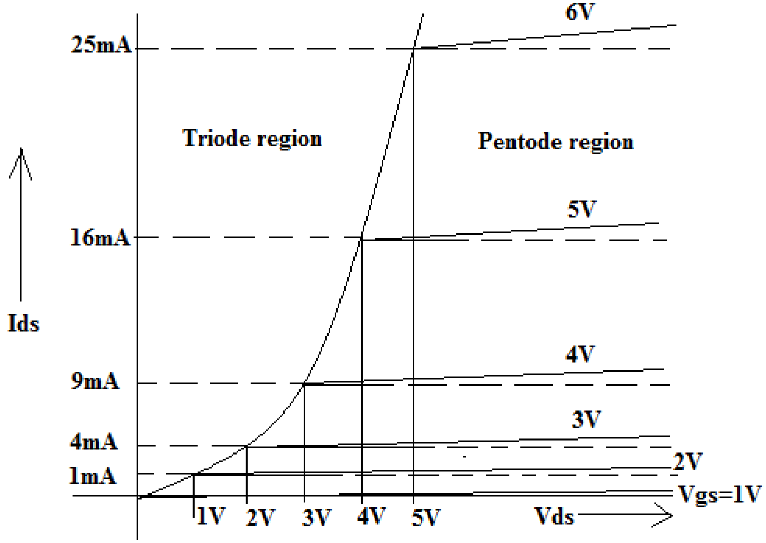


Figure 7.3.2. The effect of Channel Length Modulation on I-V family of curves of (E)NMOS.

Channel Length Modulation is included in the saturated drain current in the following manner:

$$I_{ds} = \left(\frac{W}{L} \times K\right)\frac{(V_{gs}-V_t)^2}{2} \times (1 + \lambda V_{ds})$$

7.3.28

Where λ = channel length modulation parameter which is dependent on channel length L and its typical values are:

$$0.001V^{-1} \leq$$

$$\lambda \le 0.10V^{-1}$$

.

Therefore the partial derivative of Ids with respect to Vds with gate voltage constant gives the the reciprocal of the incremental output resistance of (E)NMOS:

$$\frac{\partial I_{ds}}{\partial V_{ds}} =$$

$$\left(\frac{W}{L} \times K\right)\frac{(V_{gs}-V_t)^2}{2} \times \lambda = \frac{1}{r_{ds}}$$

7.3.29

The overall dependence of 1/rds is :

$$\frac{1}{r_{ds}} \propto \frac{1}{L^2}$$

7.3.30.

The incremental model of MOSFET incorporating transconductance and channel length modulation is given in Figure 7.3.1.

7.3.8. Figure of Merit of MOSFET.

The Unity Gain Bandwidth of MOSFET defines the Figure of Merit of MOSFET. Using Equation 7.3.27:

$$Figure\ of\ Merit = \omega_u = \frac{g_m}{C_{gs}+C_{gd}} = Unity\ Current\ Gain\ BW$$

7.3.31

Transconductance is given as:

Or

$$g_m =$$

$$\beta (V_{gs} - V_t)$$

7.3.23

Where

$$\beta = \frac{\mu_n \epsilon_{ox} \epsilon_0}{D} \times \frac{W}{L} = K \times \frac{W}{L}$$

Gate Capacitance is given as:

$$C_g = C_{ox} \times WL = \frac{\epsilon_{ox} \epsilon_0}{D} \times WL \qquad 7.3.13$$

Substituting Eq.7.3.23 and 7.3.13 in Eq.7.3.31 we get:

$$Figure\ of\ Merit = \frac{\mu_n(V_{gs} - V_t)}{L^2}$$

From Equation 7.3.7

$$\tau_{sd} = \frac{L}{v_{drift}} = \frac{L^2}{\mu_n V_{ds}} \qquad 7.3.7$$

Therefore Figure of Merit is the reciprocal of the transit time across the channel.

Larger is the electron mobility, better will be the Figure of Merit. Hence $<100>$ orientation Si Substrate is chosen for fabrication of CMOS circuis. The mobility of electron and hole is always much larger in $<100>$ orientation substrate than that in $<111>$ orientation substrate. In Table 7.3.2 a comparative study of the electron mobility in $<111>$ and $<100>$ substrate is given.

Table 7.3.2. Comparative study of mobilities in $<100>$ and $<111>$orientation substrate in thin channel and in bulk.

	Bulk mobility <100>	2D channel mobility <100>	2D Channel mobility <111>
μ_n	1250cm^2/(V-sec)	650 cm^2/(V-sec)	500 cm^2/(V-sec)
μ_p	480 cm^2/(V-sec)	240 cm^2/(V-sec)	216 cm^2/(V-sec)

For fast CMOS circuits fabrications, <100> Si Substrate is the choice of material in Industries.

SSPD_Chapter 7_Part 3_Electrical Properties of MOScontinued3

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOS coninued3 is the continuation of Electrical Properties of MOS. This describes that the proper circuit working is critically dependent on the correct geometrical dimensions of MOS transistors.

SSPD_Chapter 7_Part 3_Electrical Properties of MOScontinued3

7.3.9 MOS as an Analog Switch- Pass Transistor.

Just as electro-magnetic Relay Switches can be used to form Logic Gates. In a similar fashion, MOS can be used as analog switches to form Logic Gates. The use of Relay Switch is shown in Figure 7.3.9.1 and use of MOS is shown in Figure 7.3.9.2.

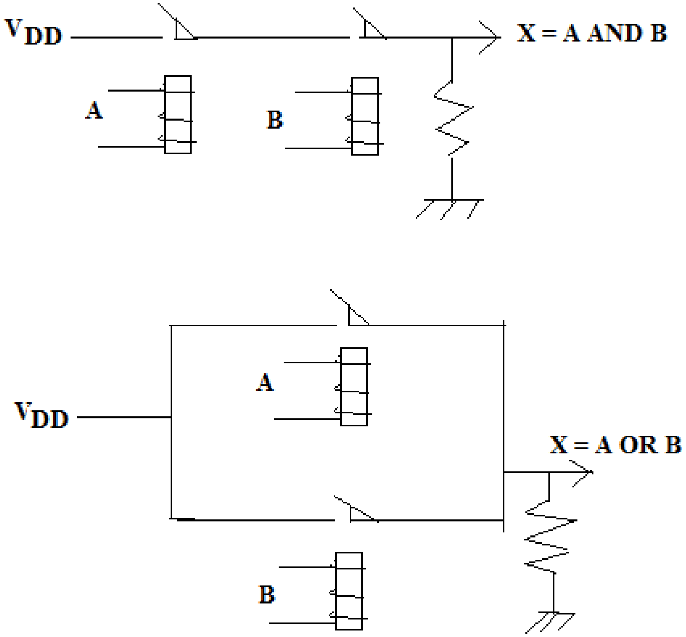


Figure 7.3.9.1. Synthesis of Logic Gates using Electro-magnetic Relays.

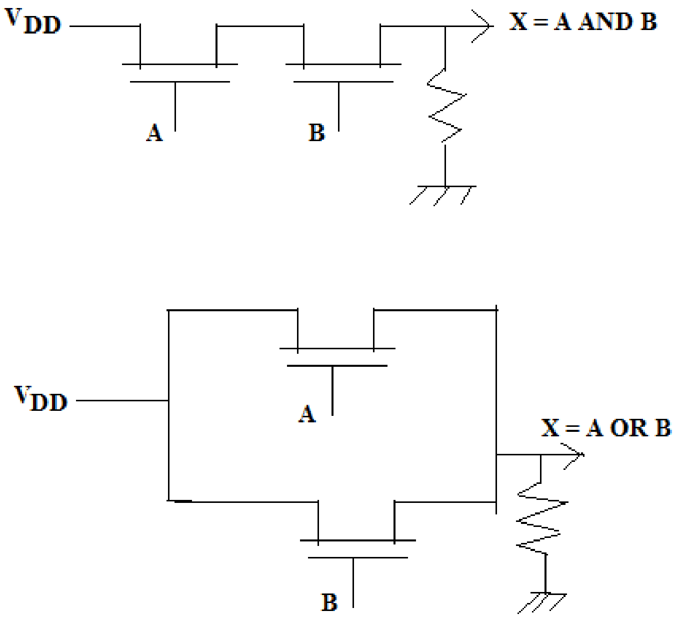


Figure 7.3.9.2 Synthesis of Logic Gates using MOS as Pass Transistors

When MOS transistors are used as analog switches they are called Pass Transistors.

Here Logic 1 is $V_{DD} - V_{th}$.

7.3.10. RTL Inverters.

RTL Inverters was the earliest building block of digital systems. It evolved to TTL gate. It further evolved to NMOS. Today CMOS Inverters have become the basic building block of most digital systems having much more superior performance characteristics as compared to the original RTL Inverter. A RTL Inverter is shown in Figure 7.3.10.1.

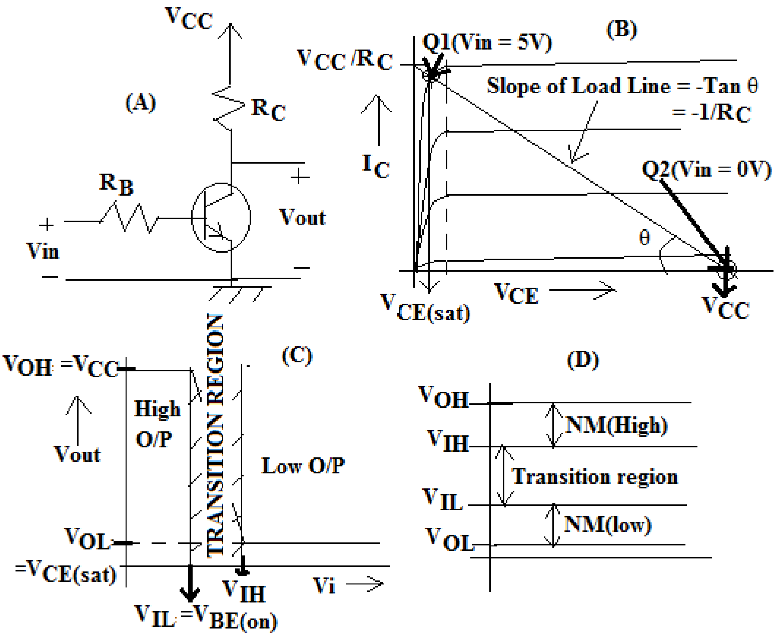


Fig.7.3.10.1 (A)RTL Circuit; (B) O/P family of curves, Load Line & Q points; (C) Transfer Characteristics (D) Definition of Noise Margin(NM)

In Figure 7.3.10.1, part (A) gives the circuit configuration. Part (B) gives the DC load line superimposed over the output family of curves. From the intersection of the DC Load Line and the output I-V curve corresponding to $V_{in} = 0V$(Low condition) and $V_{in} = 5V$(High condition) we obtain two Q points. Table 7.3.10.1 gives the two Q points.

Table 7.3.10.1 The two Q points corresponding to I/P low and I/P high.

DC operating points or Quiescent points	Vin	Vce = Vout	Ic	
Q1	5V	0.2V = Vce(sat) O/P = LOW	(Vcc)/Rc	Transistor is driven into saturation
Q2	0V	Vcc O/P = HIGH	0mA	Transistor is cut-off

Part(C) gives the transfer characteristics and Part (D) gives the noise margin. As is evident from Figure (D), noise margin can be maximized by reducing the transition region and this precisely what happens in CMOS inverter. In CMOS Inverter, transition region is 0 V and

$$NM(H) = NM(L) = \frac{V_{DD}}{2} \qquad 7.3.10.1$$

The central point of this brief review of RTL Inverter is to illustrate the principle by which Transfer Characteristics is arrived. We draw the load line over the family of output curves, determine the Q-points for the two binary states of Input and from the Q-points we arrive at the transfer characteristics of the Inverter. From the transfer characteristics we arrive at the noise margin. This is the procedure we will adopt to arrive at the performance characteristics of the subsequent inverters.

7.3.10.1. The NMOS Inverter.

MOS inverters are synthesized in the same manner as RTL Inverter in Figure 7.3.10.1. Since passive resistances are difficult to integrate as they occupy a large area on silicon die hence passive resistive loads are an anathema for ICs . The passive loads have been completely replaced by active loads. In Figure 7.3.10.2, an NMOS inverter is shown.

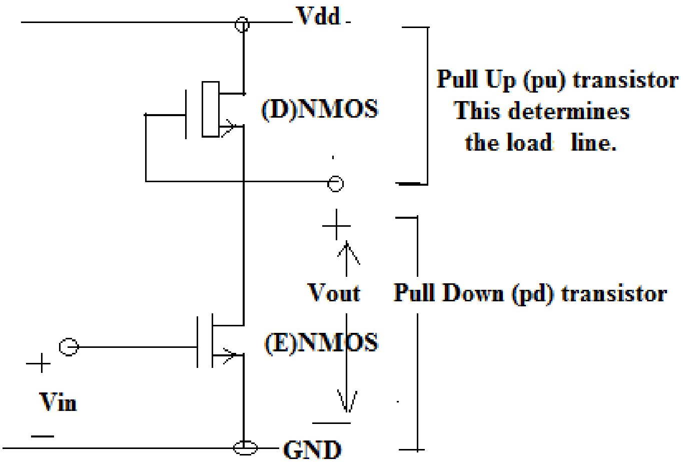


Figure 7.3.10.2. NMOS Inverter.

In the NMOS inverter, (E)NMOS acts as the inverter driver which is called the pull-down transistor and (D)NMOS acts as the active load of the driver also known as pull-up transistor. First we will examine how (D)NMOS acts as resistive active load of the logic circuit.

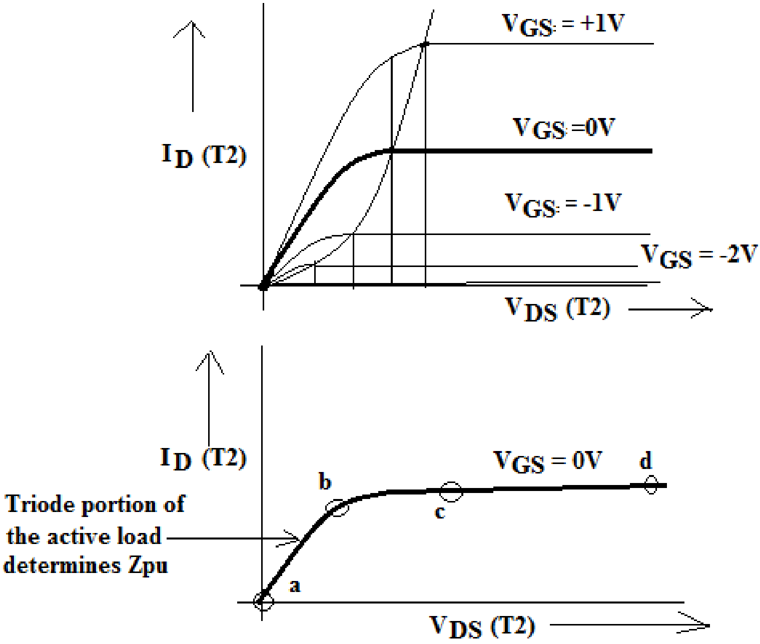


Figure 7.3.10.3. (D)NMOS source and gate are shorted together. Between DRAIN and SOURCE it acts as non-linear resistance.

(D)NMOS is transistor 2 which acts as the load to the driver transistor 1. Load (D)NMOS has its Gate shorted to Source hence $V_{GS} = 0V$ curve applies to the load I-V characteristics which is shown in the lower graph in isolation. It is a non-linear resistance accordingly the load line produced on the output plane of driver transistor (E)NMOS is also non-linear as shown in Figure 7.3.10.4. The slope of the load line is primarily determined by the Triode Region of (D)NMOS.

$$I_{ds} = \left(\frac{W}{L}\times K\right)\left[(V_{gs}-V_t)V_{ds}-\frac{V_{ds}^2}{2}\right] \tag{7.3.14}$$

$$K = \frac{\in_{ins}\in_0\ \mu_n}{D}$$

At low Vds the above equation reduces to:

$$I_{ds} = \left(\frac{W}{L}\times K\right)\left[\,(V_{gs}-V_t)V_{ds}\right]$$

Therefore:

$$\frac{1}{Z_{pu}} = \frac{W_2}{L_2}\times K\times(0-V_t) \tag{7.3.10.1.1}$$

$W\ 2\ /L\ 2\ =$ aspect ratio of the active load or transistor 2.

In Figure 7.3.10.4., the non-linear load line is drawn on the output plane of (E)NMOS and the intersection of the load line and the output curve under input high condition and under input low condition gives us the two Q points of the driver transistor as shown in Table 7.3.10.2..

Table 7.3.10.2 The two Q points corresponding to I/P low and I/P high.

Q point	Vin	Vds(T1)	Id(T1)		
(d)	+4V	Vds(LOW)	Ids*	T1 is ON	T2 is ON
(a)	0V	Vdd	0mA	T1 is OFF	T2 is ON.

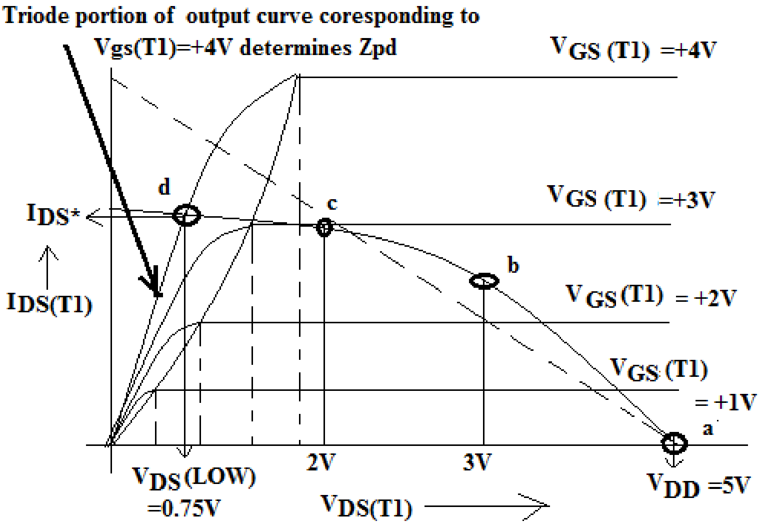


Figure 7.3.10.4. The Q points of (E)NMOS inverter under I/P low and I/P high condition.

Under I/P low, (a) is the Q point and under I/P high, (d) is the Q point

The triode portion of the output curve corresponding to $V_{GS}(T1) = +4V$ determines Z_{pd}.

From Equation 7.3.14, Z_{pd} is derived to be:

$$\frac{1}{Z_{pd}} = \frac{W_1}{L_1} \times K \times (4 - V_t) \qquad 7.3.10.1.2$$

From Figure 7.3.10.4, we can extract the inverter transfer characteristics. The point of intersections between the load line and output curves corresponding to different values of input voltage = gate voltage give the transfer characteristics of the inverter as tabulated in Table 7.3.10.3 and shown in Figure 7.3.10.5.

Table 7.3.10.3

Vin	0	2.5V	3V	4V
Vout	5V	3V	2V	0.75V

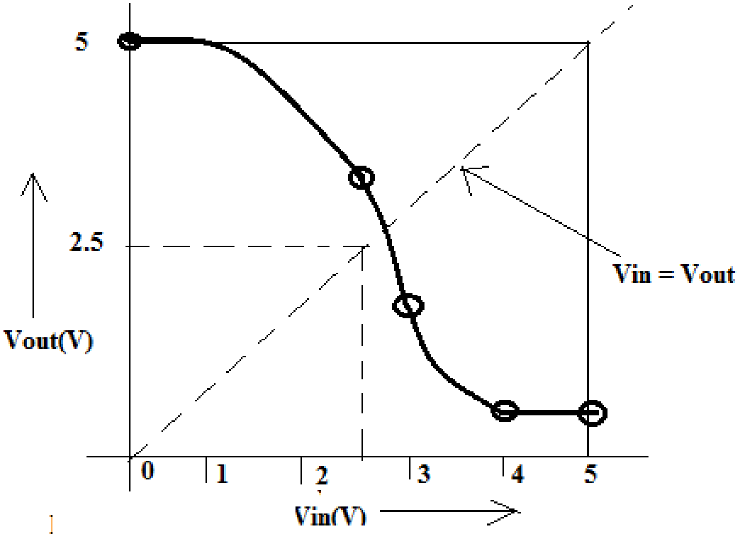


Figure 7.3.10.5. Transfer characteristics of NMOS Inverter

This transfer characteristics and inversion pont is sensitive to geometrical variations in W and L and to the variation in the ratio of pull-up to pull-down resistances ratio as shown in Figure 7.3.10.6. The point where the locus of Vin = Vout and Transfer Characteristics intersect is called the inversion point(Vinv). We have the best Noise Margin if Vinv = 0.5Vdd.

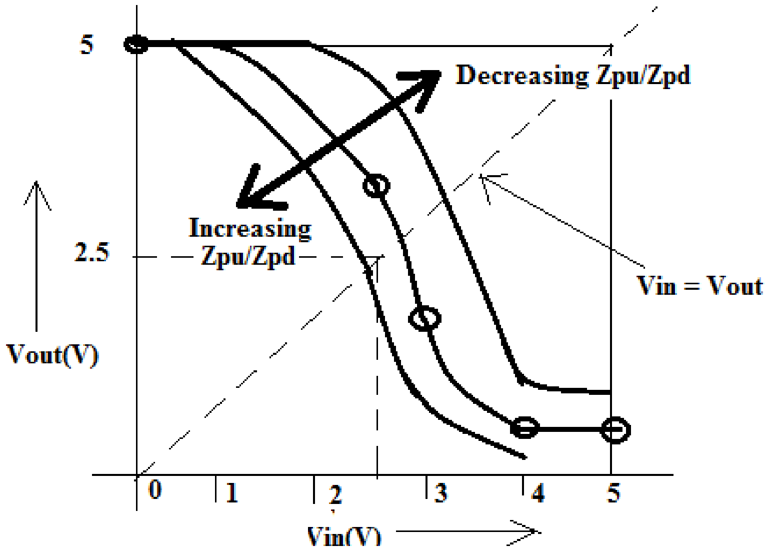


Figure 7.3.10.6. Transfer characteristics of NMOS Inverter

.

7.3.10.2. Theoretical formulation of Pull-Up to Pull-down resistances for different configurations.

There can be two configurations of NMOS Inverter. Inverter can be driven by a preceding inverter or it can be driven by a Pass Transitor as shown in Figure 7.3.10.7.

Refer to Figure 7.3.10.2. For best noise margin we choose Vin = Vinv = 0.5Vdd. At this point both the transistors are in Saturation Region(Pentode Region). In Pentode Region the following equation is applicable:

$$I_{ds} = \left(\frac{W}{L} \times K\right)\frac{(V_{gs}-V_t)^2}{2}$$

7.3.16

Under steady state condition:

IDS2 = IDS1.

Substituting the appropriate vales in Eq.7.3.16 we get:

$$I_{ds1} = \left(\frac{Wpd}{Lpd} \times K\right)\frac{(V_{inv} - V_t)^2}{2} = I_{ds2} = \left(\frac{Wpu}{Lpu} \times K\right)\frac{(0 - V_{td})^2}{2}$$

Let

$$\frac{Wpd}{Lpd} = \frac{1}{Z_{pd}}\ and\ \frac{Wpu}{Lpu} = \frac{1}{Z_{pu}}$$

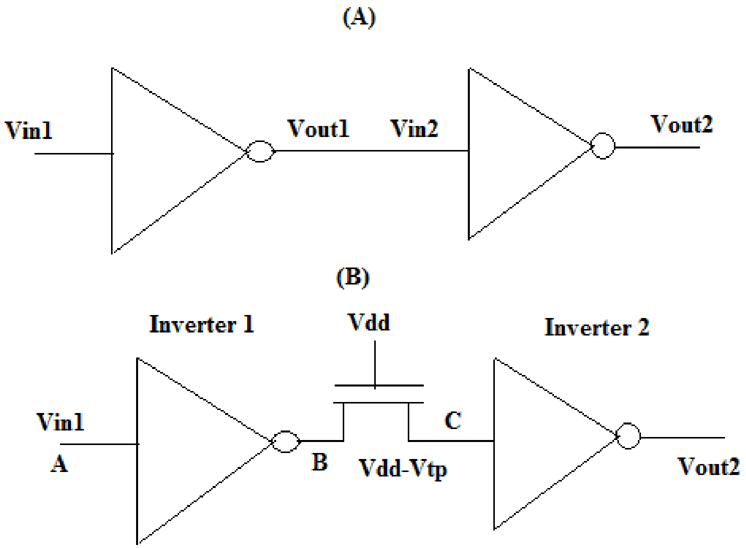


Figure 7.3.10.7. (A)Inverter driven by a preceding Inverter (B) Inverter driven by a pass transistor.

Making the proper substitutions we obtain:

$$\left(\frac{1}{Zpd}\right)\frac{(V_{inv}-V_t)^2}{1} = \left(\frac{1}{Zpu}\right)\frac{(0-V_{td})^2}{1}$$

Further Simlification yields:

$$V_{inv} = V_t - \frac{V_{td}}{\sqrt{Z_{pu}/Z_{pd}}} \tag{7.3.10.1.3}$$

By substituting the typical values of different parameters namely:

V_t = 1V, V_{td} = -3V, V_{inv} = 2.5V (for equal noise margin) we get:

$$\frac{Z_{pu}}{Z_{pd}} = \frac{4}{1}$$

7.3.10.1.4

7.3.10.2.1. Pull-Up to Pull-down ratio when NMOS Invereter is driven through pass transistors.

The second arrangement in Figure 7.3.10.7 is an NMOS Inverter driven by one or more pass transistors. Here the signal while being transmitted through Pass Transistor will get degraded (Vdd gets reduced to [Vdd-Vtp]) and may give erroneous result. The requirement is that the degradation of the signal by V_{tp} should not effect correct signal processing This implies that

Inverter 1: Vin=Vdd gives Vout=0.75V and for

Inverter 2:Vin=Vdd-Vtp should also give Vout=0.75V.

Table 7.3.10.4 gives the states of the 4 transistors in the two inverters 1 and 2 when driven by Vdd and [Vdd-Vtp] respectively.

Table 7.3.10.4. The states of the 4 transistors in

the two inverters 1 and 2 when driven by Vdd and [Vdd-Vtp] respectively.(Refer to Figure 7.3.10.4)

Inverter	Vin	pd(T1)	Vout	pu(T2)
1	Vdd	ON in Triode region	0.75V	ON in Pentode region
2	Vdd-Vtp	ON in Triode region	0.75V	ON in Pentode region

For Inverter1, in pd(T1):

$$I_{ds1} = \left(\frac{Wpd1}{Lpd1} \times K\right)\left[(V_{dd} - V_t)V_{ds1} - \frac{V_{ds1}^2}{2}\right] \qquad 7.3.10.1.5$$

$$R_1 = \frac{V_{ds1}}{I_{ds1}} = \frac{1}{K} \times Z_{pd1}\left(\frac{1}{(V_{dd} - V_t) - \frac{V_{ds1}}{2}}\right)$$

Since $V_{ds1} = 0.75V$ therefore $V_{ds1}/2$ can be neglected and the expression of R_1 is as follows:

$$R_1 = \frac{1}{K} \times Z_{pd1} (\frac{1}{(V_{dd} - V_t)})$$

For the load, VDS2 = 4.25V hence puT(2) is in Pentode region with Vgs = 0V.the transistor current I is:

$$I_{ds2} = \left(\frac{Wpu1}{Lpu1} \times K\right) \frac{(0 - V_{td})^2}{2}$$

$$I_{ds2} = \left(\frac{1}{Zpu1} \times K\right) \frac{(0 - V_{td})^2}{2}$$

Here Zpd1 and Zpu1 refer to the transistors in Inverter1. The circuit model for this I/P high is shown in Figure 7.3.10.8. pd(T1) is represented by R1 and pu(T2) is represented by a current source IDS2.

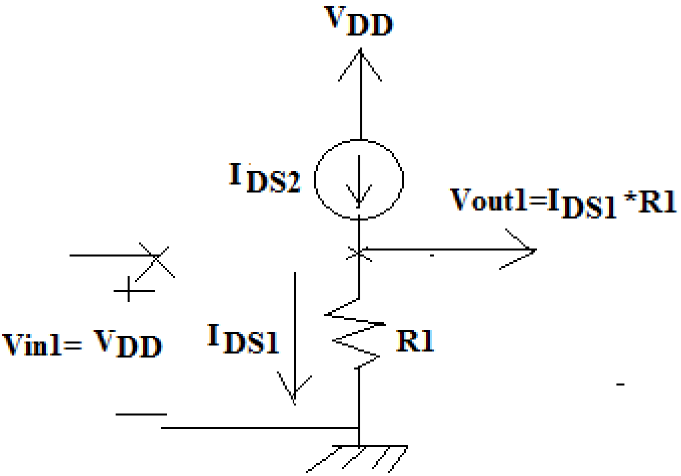


Figure 7.3.10.8. Circuit model of Inverter1 when Vin= VDD.

But $I_{DS1} = I_{DS2}$ therefore;

$$V_{out1} = R_1 \times I_{DS1} = R_1 \times I_{DS2}$$

Substituting the values of R1 and I_{DS2} from above we get:

$$V_{out1} = \frac{1}{K} \times Z_{pd1} (\frac{1}{(V_{dd} - V_t)}) \times \left(\frac{1}{Zpu1} \times K\right) \frac{(0 - V_{td})^2}{2}$$

$$V_{out1} = \frac{Z_{pd1}}{Z_{pu1}} \times \frac{V_{td}^2}{(V_{DD} - V_t)} \times \frac{1}{2} \qquad 7.3.10.1.6$$

Consider Inverter 2: when Vin = V DD -V tp .

We do exactly the same calculation as for Inverter1.

Pull-down NMOS is R_2.

Pull-up NMOS is Saturated.

$$R_2 = \frac{1}{K} \times Z_{pd2}\left(\frac{1}{(V_{dd} - V_{tp} - V_t)}\right)$$

$$I_{ds2}\big|\ inv2 = \left(\frac{1}{Zpu2} \times K\right) \frac{(V_{td})^2}{2}$$

Therefore V_{out2} of Inverter 2 with $V_{in} = V_{DD} - V_{tp}$ is given as:

$$V_{out2} = \frac{Z_{pd2}}{Z_{pu2}} \times \frac{V_{td}^2}{(V_{DD} - V_{tp} - V_t)} \times \frac{1}{2} \qquad 7.3.10.1.7$$

If the pass transistor introduces no degradation in operation then in both the inverters we should get the same output even though the second inverter is receiving a degraded Input for High condition.

Therefore

$$V_{out1} = \frac{Z_{pd1}}{Z_{pu1}} \times \frac{V_{td}^2}{(V_{DD} - V_t)} \times \frac{1}{2} = V_{out2} = \frac{Z_{pd2}}{Z_{pu2}} \times \frac{V_{td}^2}{(V_{DD} - V_{tp} - V_t)} \times \frac{1}{2}$$

Therefore

$$\frac{Z_{pu2}}{Z_{pd2}} = \frac{Z_{pu1}}{Z_{pd1}} \times \frac{(V_{DD} - V_t)}{(V_{DD} - V_{tp} - V_t)} \tag{7.3.10.1.8}$$

Substiturting the defacto values:

$V_t = 0.2V_{DD} = 1V$ and $V_{tp} = 0.3V_{DD} = 1.5V$

We get:

$$\frac{Z_{pu2}}{Z_{pd2}} = \frac{Z_{pu1}}{Z_{pd1}} \times \frac{(V_{DD} - 0.2V_{DD})}{(V_{DD} - 0.3V_{DD} - 0.2V_{DD})}$$

Or

$$\frac{Z_{pu2}}{Z_{pd2}} = \frac{Z_{pu1}}{Z_{pd1}} \times \frac{(1 - 0.2)}{(1 - 0.2 - 0.3)} = \frac{Z_{pu1}}{Z_{pd1}} \times \frac{(0.8)}{(0.5)}$$

Or

$$\frac{Z_{pu2}}{Z_{pd2}} = \frac{Z_{pu1}}{Z_{pd1}} \times 1.6 \Rightarrow \frac{Z_{pu1}}{Z_{pd1}} \times 2$$

We saw that Inverter driven directly by Inverter should have

:

$$\frac{Z_{pu}}{Z_{pd}} = 4$$

Inverter driven through a pass transistor should have:

$$\frac{Z_{pu}}{Z_{pd}} = 2 \times 4 = 8$$

$$where\ Z = \frac{L}{W}$$

SSPD_Chapter 7_Part 3_Basic Electrical Properties_continued4.

SSPD_Chapter 7_Part 3 is continued. Here we discuss the different pull up configurations we use to realize NMOS inverter. It is shown that CMOS is the best configuration from power conservation point of view though heat management remains a problem at high switching speed.

SSPD_Chapter 7_Part 3_Basic Electrical Properties_continued4.

7.3.11 Alternative forms of NAND Gates/Alternative forms of Pull-up configuration.

Just like RTL Logic, MOS Logic can have a passive load or an active load. From integration point of view we always prefer an active load in comparison to passive load.

Passive Load is a resistance and resistance occupies too much real estate on the chip as well as it has a limited options. Therefore we use an active device as the active load. In BJT Technology we have Current Mirror, Symmetrical Widlar and Widlar being used as the active load of differential amplifier which is the basic building block of an Op. Amp.

In Figure 7.3.11.1 we show the BJT Current Mirror being used as active load in differential amplifier.

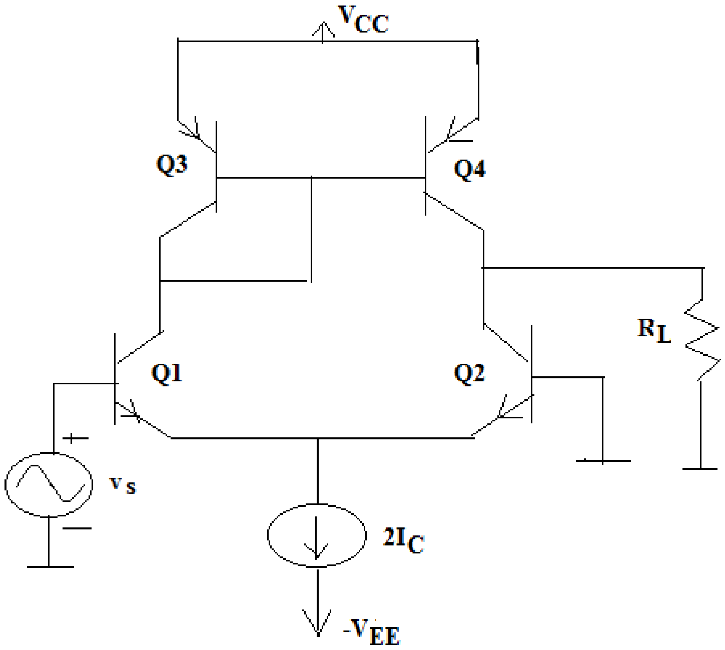


Figure 7.3.11.1. A differential amplifier using a current mirror as an active load.

Here Q3 and Q4 constitute the Current Mirror. The two together act as the active load of the differential amplifier. Q1 and Q2 are the drivers of the differential amplifier. In exactly the same manner MOS can also be used as an active load .

7.3.11.1. NMOS Inverter with (D)NMOS as pu transistor.

Passive load Pull-up configuration is technically unfeasible for IC technology. Hence we always go for Active Load Pull-Up configuration. The first configuration has already been discussed in Section

7.3.10. In this NMOS inverter with (D)NMOS as the Pull-Up transistor the biggest drawback is the standby power dissipation. When Vin = HIGH, Vout=LOW and both transistors are ON leading to rail to rail current flow hence power dissipated will be typically 5V × 5mA = 25mW.

While switching the outut from ‘1’ to ‘0’, the actual switching starts when Vin has exceeded Vt [threshold voltage of (E)NMOS].

While switching the output from ‘1’ to ‘0’, load capacitance is rapidly discharged through the pd transistor which is in triode region. Pd transistor provides low resistance path hence time constant of discharge is short.

Similarly while switching from ‘0’ to ‘1’, load capacitance rapidly charges through pu transistor. Pu transistor also is in triode region for the latter part of switching from ‘0’ to ‘1’.

This is favourable feature which was encountered in Totem Pole configuration of TTL gates. This helps improve the switching speed but the standby dissipation under Vi = HIGH disfavours this circuit hence it was left out as IC Technology progressed.

7.3.11.2.NMOS Inverter with (E)NMOS as pu transistor

Figure 7.3.11.2 describes NMOS Inverter with

(E)NMOS as pu transitor. Here also standby power dissipation is high when Vin= HIGH.

Vout never reaches V_{DD} because of voltage equal to threshold voltage dropping along the channel.

By deriving V_{GG} from a switching clock source, dissipation can be reduced. When Vin = HIGH we keep Gate of the active load LOW so that pu(T1) is turned off. So only driver (T2) is turned ON. T2 provides a low resistance path for the discharge of load capacitance when Vin = HIGH.

When Vin=LOW, T2 is OFF. During that period of Vin=LOW, we keep the gate of active load(T1) HIGH so that T1 is ON and low resistance path for charging the capacitive load. Hence Load Capacitance rapidly charges from ‘0’ to ‘1’ with a short time constant of charging.

This configuration is fast as well as it has low standby power but it has the added circuit complexity because of synchronized gate input of the active load. Hence this configuration has fallen into disfavour.

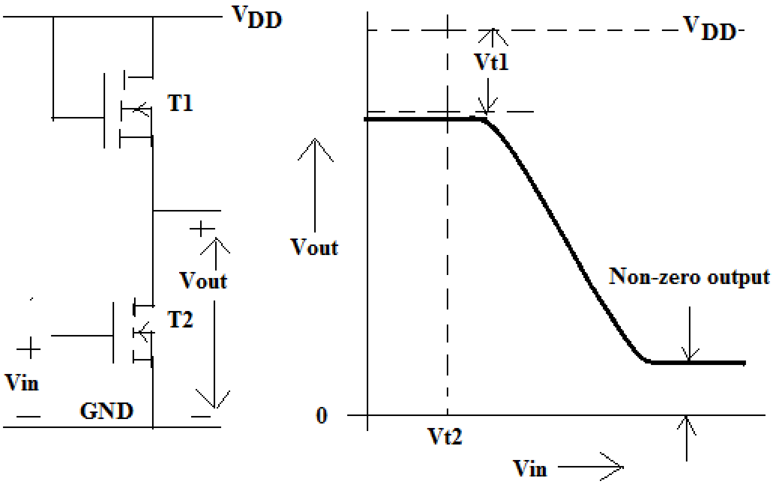


Figure 7.3.11.2. NMOS Inverter with (E)NMOS pull-up transistor.

7.3.11.2.(E)NMOS Inverter with (E)PMOS as pu transistor- complementary transitor pull-up also known as CMOS Logic.

A NMOS Inverter with a (E)PMOS as pu transistor is shown in Figure 7.3.11.3. This complementary MOS configuration is known as CMOS. The permissible states are given in Table 7.3.11.1.

Table 7.3.11.1. The two permissible states of CMOS.

		(E)NMOS	(E)PMOS	Vout
Vin	LOW	OFF	ON(sourcing	HIGH(O/

			the current to the capacitive load)	P has a low resistance path through ON PMOS to V_{DD})
Vin	HIGH	ON(sinking the current from the capacitive load)	OFF	LOW(O/P has a low resistance path through ON NMOS to the GND)

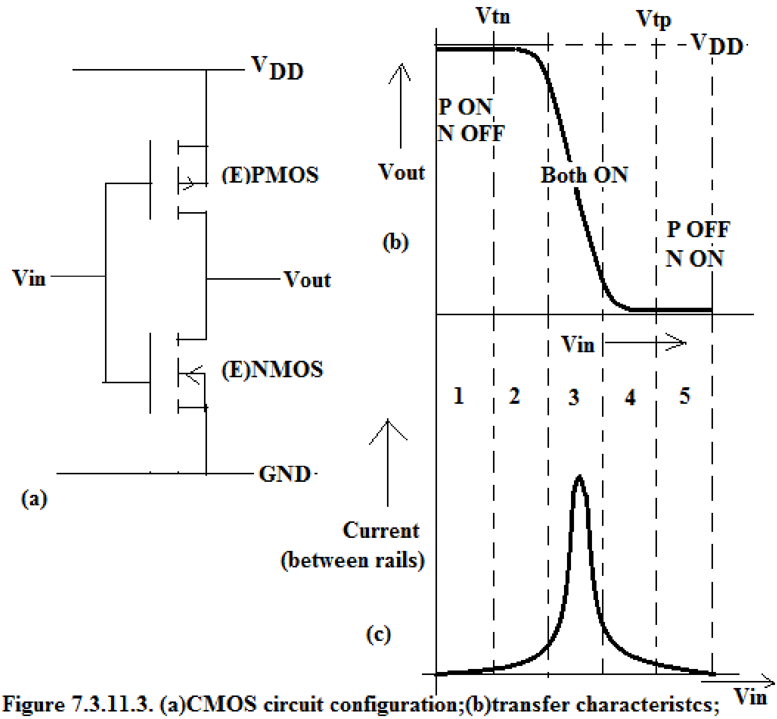


Figure 7.3.11.3. (a)CMOS circuit configuration;(b)transfer characteristcs; (c)CMOS inverter current versus Vin

Part (a) of Figure 7.3.11.3 gives the circuit configuration, Part(b) gives the transfer characteristics and Part (c) gives the current flow.

In Part(b) we see that full logical levels are realized i.e. O/P HIGH = V_{DD} and O/P LOW = 0V.

In Part (c) we see that under standby the current from rail to rail is zero. Hence standby dissipation is zero. Hence CMOS is also known as NanoWatt Logic.

During the switching there is a current flow from rail to rail hence dissipation per gate is directly proportional to clocking speed. In Figure 7.3.11.4 we compare the dissipation curves of TTL Logic and CMOS Logic under standby as well as under switching condition.

For identical geometries PMOS is slower than NMOS on account of lower mobility of holes. Hence to achieve identical switching time from LOW to HIGH and from HIGH to LOW, geometries will have to be optimized as we have to optimize for different configuration of NMOS inverter.

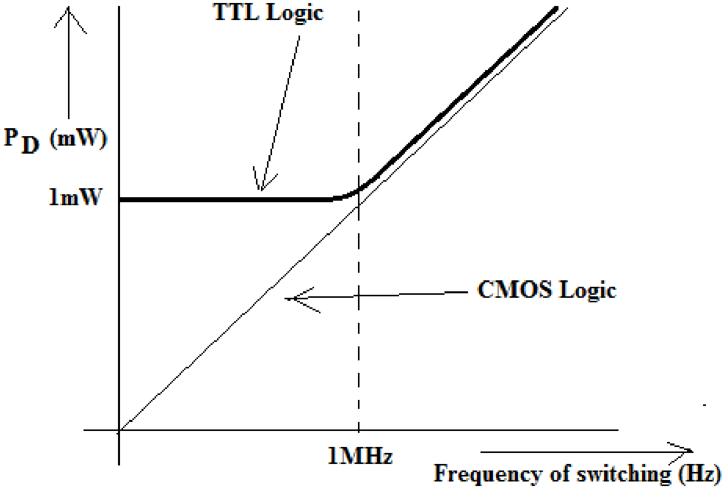


Figure 7.3.11.4. Dissipation curves of TTL Logic and CMOS Logic.

From here onward we will be focused on CMOS Logic and its analysis.

SSPD_Chapter 7_Part3_Basic Electrical Properties of MOScontinued5.

SSPD_Chapter 7_Part3_Basic Electrical Properties of MOScontinued5 describes the geometric optimization required for getting the best performance of the CMOS inverter.

SSPD_Chapter 7_Part3_Basic Electrical Properties of MOScontinued5.

7.3.12. Optimization of geometric parameters to achieve best ciruit performance of CMOS inverter.

We have seen in Section 7.3.3.1 that in triode region the drain current is:

$$I_{ds} = \left(\frac{W}{L} \times K\right)\left[\left(V_{gs} - V_t\right)V_{ds} - \frac{V_{ds}^2}{2}\right] \qquad 7.3.12.1$$

$$K = \frac{\epsilon_{ins}\epsilon_0\ \mu_n}{D} = Technology\ dependent\ parameter$$

In pentode region the drain current is:

$$I_{ds} = \left(\frac{W}{L} \times K\right)\frac{(V_{gs}-V_t)^2}{2} \qquad 7.3.12.2$$

$$\beta = \left(\frac{W}{L} \times K\right)\ and\ \frac{W}{L} = aspect\ ratio = geometric\ parameter$$

Therefore:

$$I_{ds} = (\beta)\frac{(V_{gs} - V_t)^2}{2} \qquad 7.3.12.3$$

Where

$$\beta = \left(\frac{W}{L} \times K\right) = \left(\frac{W}{L} \times \frac{\epsilon_{ins}\epsilon_0\ \mu_n}{D}\right) \qquad 7.3.12.4$$

For NMOS:

$$\beta_n = \left(\frac{\epsilon_{ins}\epsilon_0}{D}\right) \times \mu_n \frac{W_n}{L_n} \qquad 7.3.12.5$$

For PMOS:

$$\beta_p = \left(\frac{\epsilon_{ins}\epsilon_0}{D}\right) \times \mu_p \frac{W_p}{L_p} \qquad 7.3.12.6$$

Referring to Figure 7.3.11.3.(c) we see that CMOS has five regions of operation:

Region 1: I/P = LOW and O/P = HIGH, PMOS is fully ON and NMOS is fully OFF and O/P is connected to VDD through ON PMOS. Full logic ‘1’ voltage is available at the O/P.

Region 5: I/P = HIGH and O/P = LOW, PMOS is fully OFF and NMOS is fully ON and O/P is connected to GND through ON NMOS. Full logic ‘0’ voltage is available at the O/P.

Region 2: Transition region. PMOS is in triode region and NMOS is in Pentode region. Therefore:

$$(I_{DS})_p = \beta_p \left[(V_{GS} - V_{Tp})V_{DS} - \frac{V_{DS}^2}{2} \right] \qquad 7.3.12.7$$

$$(I_{DS})_n = (\beta_n)\frac{(V_{GS} - V_{Tn})^2}{2} \qquad 7.3.12.8$$

Region 4: Transition region. PMOS is in pentode region and NMOS is in trode region.

Therefore

$$(I_{DS})_n = \beta_n \left[(V_{GS} - V_{Tn})V_{DS} - \frac{V_{DS}^2}{2} \right] \qquad 7.3.12.9$$

$$(I_{DS})_p = (\beta_p)\frac{(V_{GS} - V_{Tp})^2}{2} \qquad 7.3.12.10$$

Region 3: Transtion region and both MOSs are in saturation region.

Therefore

$$(I_{DS})_n = (\beta_n)\frac{(V_{GS} - V_{Tn})^2}{2} \qquad 7.3.12.11$$

$$(I_{DS})_p = -(\beta_p)\frac{(V_{GS} - V_{Tp})^2}{2} \qquad 7.3.12.12$$

For PMOS: $V_{GS} = V_{in} - V_{DD}$ since source of PMOS is connected to positive bus.

For NMOS: $V_{GS} = V_{in}$ since source of NMOS is connected to GND bus.

Therefore

$$(I_{DS})_n = (\beta_n)\frac{(V_{in} - V_{Tn})^2}{2} \qquad 7.3.12.13$$

$$(I_{DS})_p = -(\beta_p)\frac{(V_{in} - V_{DD} - V_{Tp})^2}{2} \qquad 7.3.12.14$$

In region 3, when ON state is being transferred from one MOS to another then maximum current flows between the rails and since the two MOS are in series hence

$$(I_{DS})_n = -(I_{DS})_p$$

Therefore:

$$(\beta_n)\frac{(V_{in}-V_{Tn})^2}{2} = (\beta_p)\frac{(V_{in}-V_{DD}-V_{Tp})^2}{2}$$

Manipulating the terms we get:

$$\frac{\beta_n}{\beta_p} = \frac{(V_{in} - V_{DD} - V_{Tp})^2}{(V_{in} - V_{Tn})^2} = k^2 \qquad 7.3.12.15$$

Square root of k2 gives two solutions:

$$k = \pm \frac{(V_{in} - V_{DD} - V_{Tp})}{(V_{in} - V_{Tn})} \qquad 7.3.12.16$$

The positive solution of k gives a physically untenable result:

$$V_{in} = \frac{-V_{DD} - V_{Tp} + kV_{Tn}}{k - 1}$$

This is physically untenable because at k = 1 we get unacceptable result for V_{in}.

The negative solution of k gives an acceptable value of V_{in} :

$$V_{in} = \frac{V_{DD} + V_{Tp} + kV_{Tn}}{k + 1}$$

Or

$$V_{in} = \frac{V_{DD} + V_{Tp} + \sqrt{\frac{\beta_n}{\beta_p}} \times V_{Tn}}{\sqrt{\frac{\beta_n}{\beta_p}} + 1} \qquad 7.3.12.17$$

If

$$(\beta_n) = (\beta_p)$$

and the two transistors are of equal magnitude threshold voltage that is:

$$V_{Tp} = -V_{Tn}$$

then

$$V_{in} = \frac{V_{DD}}{2}$$

This implies that the switch over is symmetrically disposed about the point :

$$V_{in} = V_{out} = \frac{V_{DD}}{2}$$

To achieve a symmetrical transfer curve it is necessary that

$$(\beta_n) = (\beta_p)$$

And

$$V_{Tp} = -V_{Tn}$$

This implies that

$$\mu_n \frac{W_n}{L_n} = \mu_p \frac{W_p}{L_p}$$

Or

$$\frac{W_p}{L_p} = \frac{\mu_n}{\mu_p} \times \frac{W_n}{L_n} = 2.5 \times \frac{W_n}{L_n}$$

As discussed in the preceding section, inorder to get optimum performance in switching speed as well as in transfer curve PMOS will have to be wider than NMOS.

In VLSI Technology, as the dimensions are being scaled short channel effects are showing up. One of the short channel effects is electric drift mobility dependence on longitudinal and transverse electric field.

Reference [“The dependence of the electron mobility on the longitudinal electric field in MOSFETs” by J.B.Rolden et al,Semiconductor

Science Technology, Vol 12,(1997),321-330] gives the following relationship:

$$\mu(E_{\parallel}, E_{\perp}) = \frac{\mu_0(E_{\perp})}{\sqrt[M]{1 + (\frac{\mu_0(E_{\perp}) \times E_{\parallel}}{v_{sat}})^M}} \qquad 7.3.12.18$$

$$E_{\parallel} = longitudinal\ electric\ field$$

$$E_{\perp} = transverse\ electric\ field.$$

$$\mu_0 = low\ field\ mobility$$

Pucknell et. al. gives the following expression:

$$\mu = \mu_Z(1 - \varphi(V_{gs} - V_t))^{-1} \qquad 7.3.12.19$$

Where φ = constant of the range 0.05;

V_t = threshold voltage including the body effect;

And μ_Z = drift mobility with zero transverse electric field.

V_{GS}/D = decides the transverse electric field;

D = gate oxide thickness.

Since mobility has transverse electric field dependence hence only at the symmetry point, k = 1 hold good.

As can be seen from Eq.7.3.12.17, at $V_{out} = V_{DD}/2$, $V_{in} = V_{DD}/2$ only at k = 1.

$$V_{in} = V_{out} = \frac{V_{DD}}{2}$$

The relation ship is :

$$\frac{W_p}{L_p} = 2.5 \times \frac{W_n}{L_n}$$

is satisfied.

If due to manufacturing tolerance k < 1 then according to Eq.7.3.12.17,

$$V_{in} > V_{out} = \frac{V_{DD}}{2}$$

If k > 1

$$V_{in} < V_{out} = \frac{V_{DD}}{2}$$

We get a similar kind of lateral shift in the transfer characteristics as shown in Figure 7.3.12.1.

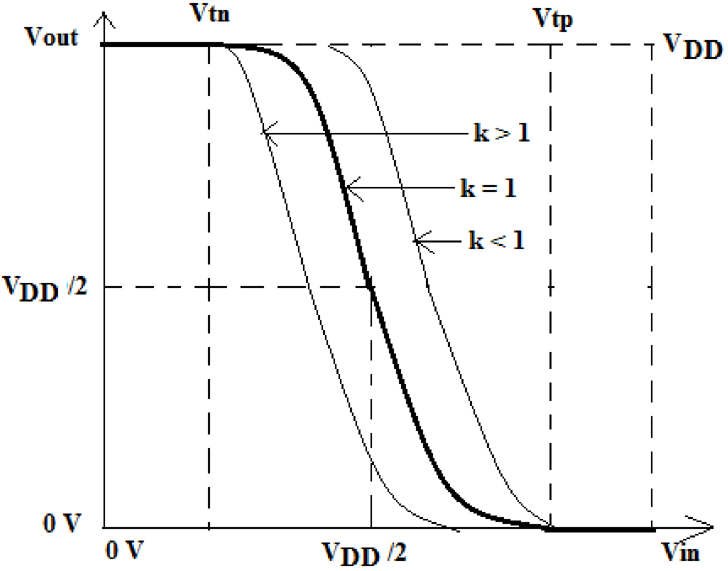


Figure 7.3.12.1. Trends in transfer characteristics with beta ratio (k)

As can be recognized that the symmetry point is:

$$V_{in} = V_{out} = \frac{V_{DD}}{2}$$

is inherently a unstable point hence the changeover from HIGH to LOW or vica versa is rapid and only

$$V_{in} = High\ and\ V_{out} = Low$$

Or

$$V_{in} = Low\ and\ V_{out} = High$$

are the two stable operating points.

Now we will look at the circuit model of CMOS.

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOScontinued6

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOScontinued6 describes the various capacitance associated NMOS in triode region, pentode region and in cut-off region.

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFETcontinued 6

7.3.13. MOS Transistor’s parasitic capacitances.

Just as parasitic capacitances Cμ and Cπ limit the frequency reponse of BJT in exactly the identical manner gate oxide capacitances limit the frequency response of MOSFET.

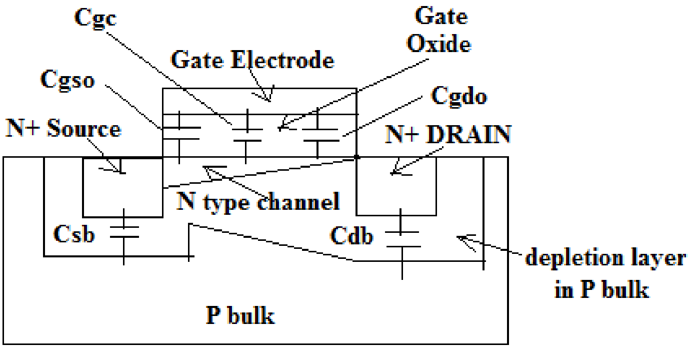


Figure 7.3.13.1. NMOS Transistor Model.

Referring to Figure 7.3.13.1. we identify the following capacitances:

CGC = gate to channel capacitance,

CGSO = gate to source overlap capacitance and

CGDO = gate to drain overlap capacitance.

Overlap capacitances occur mainly due to:

1. Gate misalignment with respect to Source and Drain.
2. Fringing field between Gate and Source/Gate and Drain.

Because of self aligned processing, CGSO and CGDO are minimized. CGC has the dominant contribution in limiting the frequency response.

Typical value of channel-gate capacitance is:

$$C_{GC} = WL \times C''_{ox} = WL \times \frac{\epsilon_{SiO2}\epsilon_0}{D_{ox}} = WL \times \frac{3.9 \times 8.854 \times 10^{-14} F/cm}{25 \times 10^{-9} m \times 100 cm}$$
$$= WL \times 0.138 \mu F/cm^2$$

7.3.13.1. Capacitance in Triode Region.

We effectively have CGS and CGD in Triode region of MOSFET. The contribution of CGC is included in these two capacitances.

$$C_{GS} = \frac{C_{GC}}{2} + C_{GSO} W \qquad 7.3.13.1$$

$$C_{GD} = \frac{C_{GC}}{2} + C_{GDO}W \quad 7.3.13.2$$

Note that COX” is Farad per unit area whereas overlap capacitance is Farad per unit length.

As seen in the Figure 7.3.13.1, in addition to the above capacitances , there is Source to Bulk capacitance (CSB) and Drain to Bulk capacitance (CDB) due to the depletion layer surrounding the Source and Drain. The reverse biased junction exists at the bottom of the source and drain. It also exists at the periphery of the source and the drain.

Source-to-Bulk capacitance =

$$C_{SB} = C_j A_S + C_{jSW} P_S \quad 7.3.13.3$$

$$C_{DB} = C_j A_D + C_{jSW} P_D \quad 7.3.13.4$$

PS and PD are the perimeters of Source and Drain.

AS and AD are the bottom area of Source and Drain.

Cj and CjSW are dependent on the reverse voltage existing between the Source(or Drain) and the bulk because junction capacitance is proportional to CS area/ depletion width and depletion width varies the square root of the reverse voltage in step kunction and it depends on the cube root of the reverse voltage in linearly graded junction.

7.3.13.2. Capacitances in Pentode or Saturation Region.

In saturation region, there is no channel on the drain side hence Gate-Channel capacitance has no contribution to Gate-Drain Capacitance. Therefore:

$$C_{GS} = \frac{2C_{GC}}{3} + C_{GSO}W \qquad 7.3.13.5$$

$$C_{GD} = 0 + C_{GDO}W \qquad 7.3.13.6$$

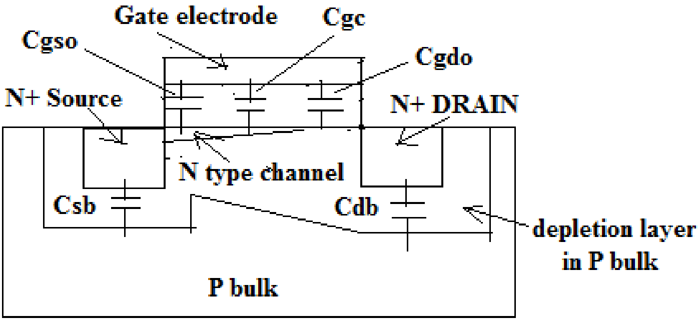


Figure 7.3.13. 2. NMOS in saturation region. The n-channel gets pinched off earlier than N+ Drain

7.3.13.2. Capacitances in Cut-off Region.

Figure 7.3.13.3 gives the cross-sectional area of NMOS which is below the threshold region hence it is in cut-off region. In cut-off region there is no channel hence no gate-channel capacitance.

$$C_{GS} = 0 + C_{GSO}W \qquad 7.3.13.7$$

$$C_{GD} = 0 + C_{GDO}W \qquad 7.3.13.8$$

$$C_{GB} = C_{GBO}W \qquad 7.3.13.9$$

CGBO = Gate-Bulk Capacitance per unit width.

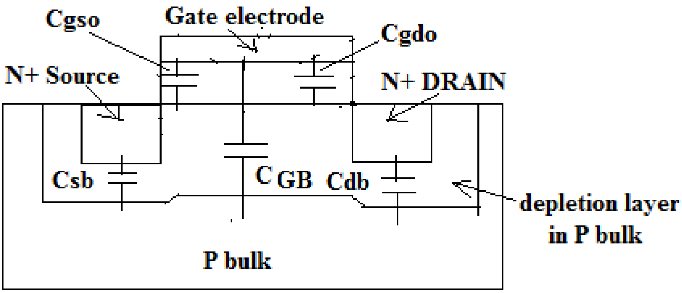


Figure 7.3.13. 3. NMOS in cut-off region that is gate voltage is below Threshold Voltage. There is no channel. The depletion region below the gate-oxide creates a small capacitance Cgb.

Generally in Digital Systems we will encounter Saturation and Cut-off conditions hence capacitance s showed in Figure 7.3.13.2 and in Figure 7.3.13.3 will be relevant to us in course of analysis.

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_continued7

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_continued7 gives the differences in the key parameters of BJT and MOSFET.

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_continued7

7.3.14. Comparative aspects of Key Parameters of CMOS and BJT.

In the Table 7.3.14.1 we enumerate the major differences of CMOS and BJT in qualitative terms then we will deal with it in quantitative terms.

Table 7.3.14.1. Comparative study of CMOS and BJT.

	Bipolar Junction Transistors	CMOS
Carriers involved in transistor action	Bipolar(both minority and majority cariers are involved)	Unipolar (only majority carriers participate)
Input dynamic	Exponential	Quadratic Device

range	device hence input dynamic range for linear applications is 5mV	hence input dynamic range for linear applications is 400mV
Noise Figure	Much higher hence not suitable for Low Noise Amplifier	At cryogenic temperatures its Noise Figure is nearly Unity hence MESFET is suitable for LNA
Packing Density	Because of isolation diffusion BJT ICs are traditionally 1-G behind in level of integration.	This has far ahead in packing density.
Speed	Historically it is much faster because transit through base width limits the switching speed. It has reached more than 100GHz	Here the transit through the channel length limits the speed. The maximum switching speed is 2GHz or more.
Transconductance	40mS at 1mA	1mS at 1mA
Drive capability(A/cm^2)	10 times more than that of CMOSHence at	

the driver stage
we use BiCMOS

7.3.14.1. Quantitative analysis of the key parameters of BJT and CMOS.

In the Pentode region of operation:

$$I_{ds} = \left(\frac{W}{L} \times K\right)\frac{(V_{gs}-V_t)^2}{2} \qquad 7.3.12.2$$

$$\beta = \left(\frac{W}{L} \times K\right) \ and \ \frac{W}{L} = aspect\ ratio = geometric\ parameter$$

Therefore:

$$I_{ds} = (\beta)\frac{(V_{gs}-V_t)^2}{2} \qquad 7.3.12.3$$

Where

$$\beta = \left(\frac{W}{L} \times K\right) = \left(\frac{W}{L} \times \frac{\epsilon_{ins}\epsilon_0\ \mu_n}{D}\right) \qquad 7.3.12.4$$

In circuit application books β is referred to as Kn , transcoductance parameter for NMNOS.

Substituting the values of the different parameters in Eq.7.3.12.4 we get:

$$C''_{ox} = \frac{\epsilon_{Sio2}\epsilon_0}{D_{ox}} = \frac{3.9 \times 8.854 \times 10^{-14}\,F/cm}{25 \times 10^{-9}m \times 100cm} = 0.138\mu F/cm^2$$

$$C''_{ox} \times \mu_n = 500\,\frac{cm^2}{V - sec} \times 0.138\,\frac{\mu F}{cm^2} = 69\,\frac{\mu A}{V^2} = K$$

Therefore β = (W/L) × 69(μA/V2)

In Table 7.3.14.2 we get the special transconductance parameter for different aspect ratio.

Table 7.3.14.2. Special Transconductance Parameter for different geometries.

W	L	W/L	β(μA/V2)
20μm	1 μm	20	1380
60 μm	3 μm	20	1380
10 μm	0.25 μm	40	2760

From Section 7.3.5 we get:

$$g_m = \beta\left(V_{gs} - V_t\right)$$

7.3.23

Where

$$\beta = \frac{\mu_n \epsilon_{ox} \epsilon_0}{D} \times \frac{W}{L} = K \times \frac{W}{L}$$

Dividing Eq.7.3.23 by 7.3.12.3 we get:

$$\frac{g_m}{I_{DS}} = \frac{2}{(V_{GS} - V_{Tn})} \qquad 7.3.14.1$$

We will choose W/L = 10μm/0.25μm = 40.

Special transconductance parameter = β = 40 × 69μA/V2 = 2760 μA/V2

Therefore:

$$If \qquad I_{ds} = (\beta)\frac{(V_{gs} - V_t)^2}{2} = 1mA$$

Then VGS-VTn =0.851V.

Substituting this value in Eq.7.3.14.1

We get :

$$\frac{g_m}{I_{DS}} = \frac{2}{(V_{GS} - V_{Tn})} = \frac{2}{0.851} = 2.35\left(\frac{1}{V}\right) \qquad 7.3.14.2$$

Similarly if we analyze BJT we get :

$$g_m = \frac{I_C}{V_t}$$

$$\frac{g_m}{I_C} = \frac{1}{V_t} \qquad 7.3.14.3$$

I$_C$ = d.c. collector current and V$_t$ = thermal voltage = kT/q = 25mV at 300K_Room Temperature.

$$\frac{g_m}{I_c} = \frac{1}{V_t} = \frac{1}{25mV} = 40(\frac{1}{V}) \qquad 7.3.14.4$$

By comparing Eq 7.3.14.2 aand 7.3.14.4 we see that for same bias current 1mA, BJT delivers 17 times transconductace as compared to that of MOSFET.

7.3.14.2. Current Delivery Capability of BJT and MOSFET.

Rewriting Eq.7.3.12.3

$$I_{ds} = (\beta)\frac{(V_{gs} - V_t)^2}{2} \qquad 7.3.12.3$$

Where

$$\beta = \frac{\mu_n \epsilon_{ox} \epsilon_0}{D} \times \frac{W}{L}$$

Therefore:

$$I_{ds} = \frac{\mu_n \epsilon_{ox} \epsilon_0}{D} \times \frac{W}{L} \times \frac{(V_{gs} - V_t)^2}{2}$$

Or

$$\frac{I_{DS}}{W \times L} = \frac{I_{DS}}{A} = \frac{\mu_n \epsilon_{ox} \epsilon_0}{D} \times \frac{1}{L^2} \times \frac{(V_{gs} - V_t)^2}{2} \qquad 7.3.14.5$$

Calculating for W/L = 10µm/0.25µm = 40. And D = 25nm we get:

$$\frac{I_{DS}}{A} = 69 \frac{\mu A}{V^2} \times \frac{1}{2 \times (250nm)^2} \times (V_{GS} - V_{Tn})^2 \qquad 7.3.14.6$$

Analyzing for IDS = 1mA we get:

$$\begin{aligned} \frac{I_{DS}}{A} &= 69 \times \frac{1}{2 \times (250 \times 10^{-9})^2} \times (0.851)^2 \frac{\mu A}{m^2} \\ &= 4 \times 10^{14} \frac{\mu A}{m^2} \qquad 7.3.14.7 \end{aligned}$$

Expressing it in A/cm2 we get:

$$\frac{I_{DS}}{A} = 40{,}000 \ \frac{A}{cm^2} \qquad 7.3.14.8$$

7.3.14.2.1. Current delivery capability of BJT.

From the Physics of BJT we have the following relationship:

$$I_C = \frac{minority\ charge\ stored\ in\ BASE}{transit\ time\ through\ the\ BASE} = \frac{Q_B}{\tau_B}$$

$$= \frac{A_E W_B \times q(\frac{n_{po}\lambda}{2})}{\tau_B} \qquad 7.3.14.9$$

Therefore:

$$\frac{I_C}{A_E} = \frac{W_B \times q \times n_{po} \times \lambda}{2\tau_B} \qquad 7.3.14.10$$

Where W_B = effective base width, n_{po} = thermal equilibrium value of minority carrier in Base; λ = Exp(V_{BE}/V_t).

From Device Physics we also know that the Collector Current is diffusion current given by the following expression:

$$I_C = \frac{-q \times A_{E\times} D_n \times gradient\ of\ minority\ charge\ \ in\ BASE}{1}$$

$$= -q \times A_{E\times} D_n \times \frac{n_{po} \times \lambda}{W_B} \qquad 7.3.14.11.$$

Since our reference current is I_C = 1mA and we choose $A_E = 1(\mu m)^2$

Therefore

$$\frac{I_C}{A_E} = \frac{1mA}{1\mu m^2} = \frac{1\times10^{-3}A}{(1\times10^{-4}cm)^2} = 10^5\frac{A}{cm^2} \qquad 7.3.14.11$$

Therefore

$$\frac{I_C}{A_E} = \left| \, q \times D_n \times \frac{n_{po} \times \lambda}{W_B} \right|$$

Therefore

$$\lambda = \frac{I_C}{A_E} \times \frac{W_B}{q \times D_n \times n_{po}} = \frac{I_C}{A_E} \times \frac{W_B}{q \times \mu_n \times V_{thermal} \times n_{po}} \qquad 7.3.14.12$$

Following are the various parameter values:q = 1.6×10^{-19}Coulombs, μ_n = bulk electron mobility = 1450 cm^2/(V-sec), Thermal Voltage = $V_{thermal}$ = kT/q = 25mV at 300Kelvin, n_{po} = thermal equilibrium value of minority carriers in Base in a NPN transistor = 10^3/cm^3 and W_B = effective base width = 1µm = 1×10^{-4}cm.

Substituing the above values in Eq.7.3.14.12 we get:

$$\lambda = 10^5 \frac{A}{cm^2} \times \frac{1 \times 10^{-4} cm}{1.6 \times 10^{-19} Coul \times 1450 \frac{cm^2}{V - s} \times 25 \times 10^{-3} V \times 10^3 cm^{-3}}$$

By inspection we see that it is a dimensionless quantity therefore:

$$\lambda = 10^5 \times \frac{1 \times 10^{-4}}{1.6 \times 10^{-19} \times 1450 \times 25 \times 10^{-3} \times 10^3} = 1.724 \times 10^{15}$$

We substitute the value of $\lambda = 1.724 \times 10^{15}$ in Eq.7.3.14.10:

$$\frac{I_C}{A_E} = \frac{W_B \times q \times n_{po} \times \lambda}{2\tau_B} = \frac{1 \times 10^{-4} cm \times 1.6 \times 10^{-19} Coul \times 10^{3} cm^{-3} \times 1.724 \times 10^{15}}{2 \times 20 \times 10^{-12} sec} \qquad 7.3.14.13$$

Eq.7.3.14.13 is dimensionally balanced.

$$\frac{I_C}{A_E} = \frac{1 \times 10^{-4} \times 1.6 \times 10^{-19} \times 10^{3} \times 1.724 \times 10^{15}}{2 \times 20 \times 10^{-12}} \frac{A}{cm^2} = 431000 \frac{A}{cm^2} \qquad 7.3.14.14$$

Comparing Eq 7.3.14.8 and Eq.7.3.14.14 we see that BJT has 10 times drive capability.

7.3.14.3.Quantitative differences in the key parameters

	BIPOLAR TRANSISTOR	MOSFET
Current drive capability for 1mA in units of A/cm2	4,31,000	40,000
Transconductance	40	2.35

Bias Current(1/V)		
Transconductace(Siemens)	Varies directly as bias current	Varies as the square root of the bias current
Transconductance as a function of geometry	Weak function	Strongly dependent on aspect ratio and oxide thickness
Transconductance as a function of process technology	Independent	Dependent through oxide thickness

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_continued8

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_continued 8 gives the configuration of BiCMOS inverter and its optimized version.

SSPD_Chapter 7_Part 3_Basic Electric Properties of MOSFET_Continued_8

7.3.15. BiCMOS Inverters.

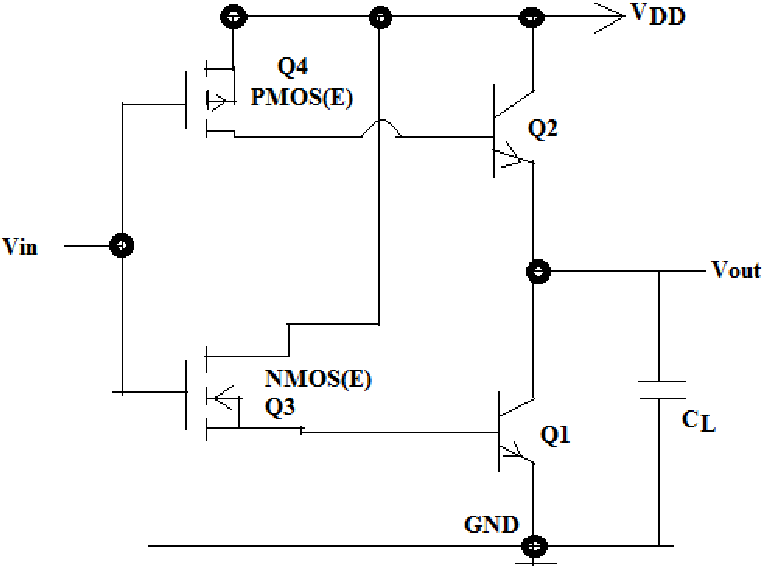


Figure 7.3.15.1. A simple BiCMOS inverter.

In the output stage where loads have to be driven, CMOS Logic Gates themselves cannot give adequate current for drivng the loads. In these circuits buffer

BJT are added as shown in Figure 7.3.15.1. The inverted logic operation is carried out as shown in Table 7.3.15.1.

Table 7.3.15.1. The Inverted Logic Operation of BiCMOS inverter.

Input	Q1	Q2	Q3	Q4	Vout
LOW(GND)	off	On and Q2 sources current to charge C_L to 5V through a low impedance path provided by on BJT Q2	off	on	HIGH = 5-V_{BE}(on)
HIGH(5V)	On and Q1 sinks current thereby discharging	off	on	off	LOW = V_{CE}(sat)

C_L through low impedance path provided by on BJT Q1

The characteristics of BiCMOS Inverter.

i.Logic levels are close to bus voltages;

ii.High input impedance;

iii.Low output impedance;

iv. It has high current drive capability and occupies a relatively small area;

v. It has large Noise Margin.

But this configuration has high standby dissipation under I/P = HIGH condition since Q1 and Q3 provide a low impedance path from Positive Bus to GND bus.

Also this configuration has no path for turning off the BJT when required.This slows the switching operation. Hence we go for improved circuit configuration.

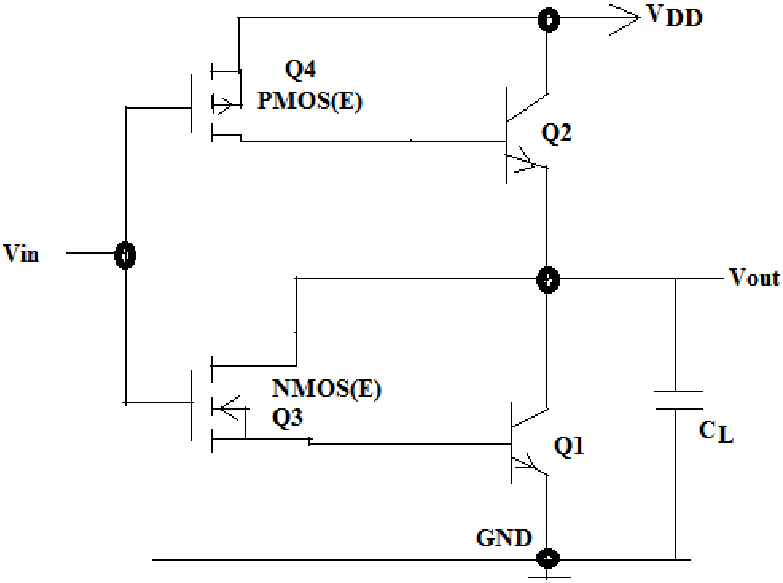


Figure 7.3.15.2. An alternative BiCMOS inverter with no static current drain under input HIGH condition.

In Figure 7.3.15.2 we have removed the direct dc path from Positive Bus to GND through Q3 and Q1 but when Vin = HIGH, Vout ≠ V_{CE}(sat) which is 0.2V but Vout = V_{BE}(on) = 0.7V. Hence output swing is reduced from full 5V to about 4V.

Also BJTs still donot have a discharge path for their base currents during turn off of the respective BJT. So a third configuration is given in Figure 7.3.15.3.

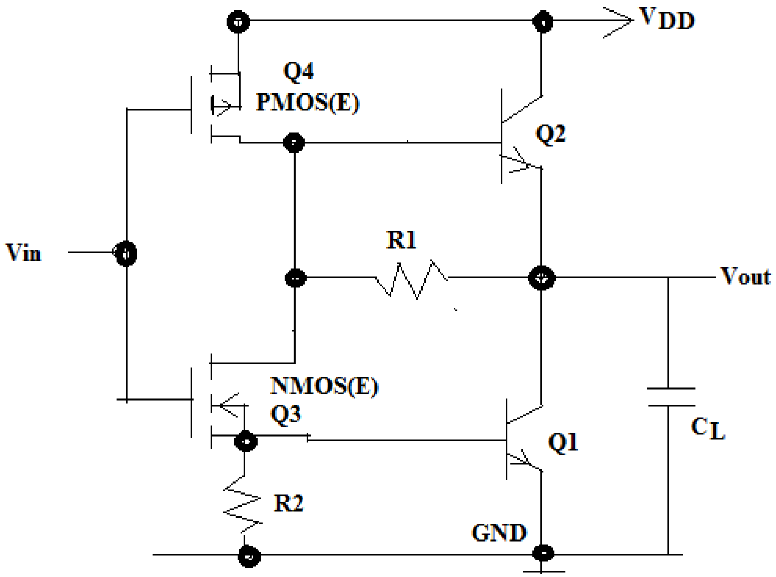


Figure 7.3.15.3. An improved BiCMOS with better O/P logic levels and with discharge paths for BJT base currents.

In the improved version shown in Figure 7.3.15.3. two resistances R1 and R2 are introduced in the discharge paths of the bases of the BJTs , Q1 and Q2. Because of the resistances, the logic levels are improved and hence voltage swing is optimized. The resistances also speed up the switching of the transistors.

Though this is an improved version of BiCMOS Inverter from IC technology point of view it is impractical because passive chip resistances occupy large real estate area and suitable values are difficult to realize. Hence chip resistances are replaced by MOS transistors as shown in Figure 7.3.15.4.

In Figure 7.3.15.4, the function of R1 is fulfilled by Q5 and function of R2 is fulfilled by Q6. The states of different transistors during Inverted Logic Operation are given in Table 7.3.15.2.

Table 7.3.15.2. Inverted Logic Operation of the optimized version ogf BiCMOS.

Vin	Q1	Q2	Q3	Q4	Q5	Q6	Vout
HIGH	SAT	OFF	ON	OFF	ON This provides the discharge path for the base current of Q2 which was ON previously	OFF	LOW = V_{CE}(s
LOW	OFF	ON	OFF	ON	OFF	ON This provides the discharge	HIGH

path for the base current of Q1 which was in SAT previously

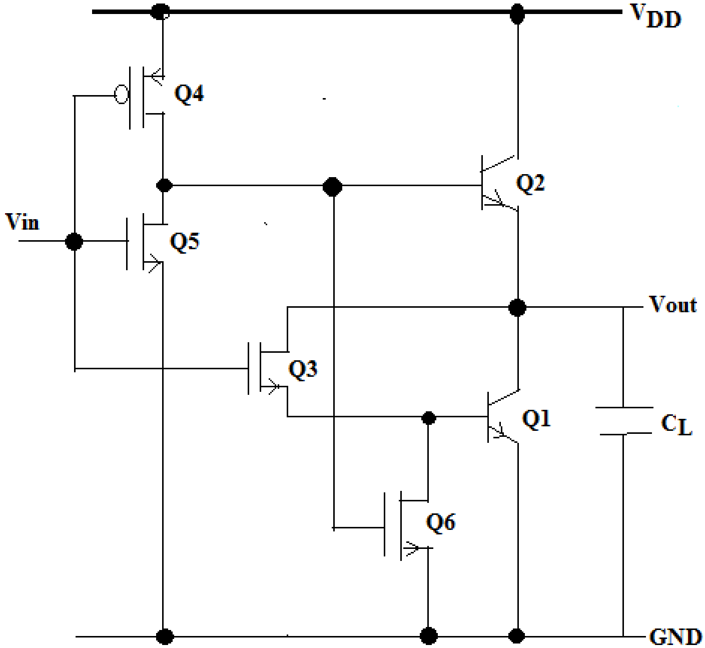


Figure 7.3.15.4. A version of BiCMOS Inverter where resistances R1 and R2 of Figure 7.3.15.3 have been replaced by MOS Transistors Q5 and Q6.

This circuit offers enormous advantages where high load current sourcing and sinking are required.

Whenever Video Monitors or Speakers are to be driven there we require BiCMOS interface at the output.

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_concluded.
Chapter 7_Part 3_Conclusion module concludes the basic electrical properties of MOSFET by discussing the latch-up problem and its remedies.

SSPD_Chapter 7_Part 3_Basic Electrical Properties of MOSFET_concluded.

7.3.16. Latch-Up problem in CMOS and its remedy.

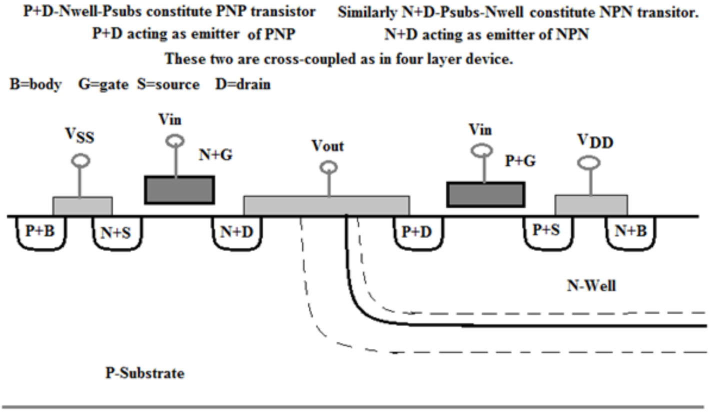
[This part is adapted from the lecture series of Steve McGarry,Department of Electronics, University of Carleton, Ottawa. e-mail:smcgarry@doe.carleton.ca]

CMOS logic is fabricated either by P-Well or N-Well Process or by Twin-Well Process. In any of these processes we have a large number of junctions as . In Figure 7.3.16.1. we give the cross-sectional view of CMOS fabricated by N-Well process. These multiple junctions give rise to three-junction four-layer device which behaves like a SCR. This we call parasitic SCR.

This arises due to cross coupling of P+NP and N+PN transistors. How these parasitic transistors action arise is shown in Figure 7.3.16.2.

If this parasitic SCR is turned ON by the glitches on the power bus or by irradiation, then a low resistance conducting path is created between

Positive Voltage Bus (V_{DD}) and GND Bus (V_{SS}). This causes irreversible damage to the IC chip. This precisely is SCR Latch-Up problem in CMOS. By careful control during fabrication, the parasitic SCR can be prevented from latching up under all conditions.


P+D-Nwell-Psubs constitute PNP transistor Similarly N+D-Psubs-Nwell constitute NPN transitor.

P+D acting as emitter of PNP N+D acting as emitter of NPN

These two are cross-coupled as in four layer device.

B=body G=gate S=source D=drain

Figure 7.3.16.1. A CMOS structure fabricatewd using N-Well in P-substrate.

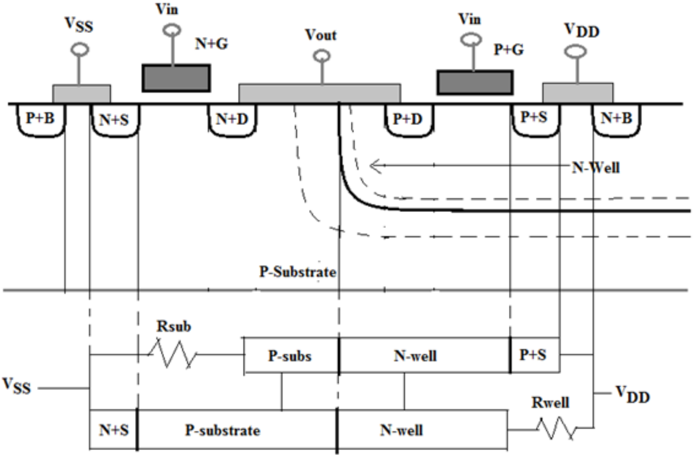


Figure 7.3.16.2. Four layers N+S, P-Substrate, N-Well and P+S constitute the parasitic SCR device.

In MOS structures the source is always shorted to the body. As seen in the above Figures, Source of NMOS i.e. N+S is connected to P+body by a metal interconnect and P+body makes an ohmic contact to P-substrate. Similarly Source of PMOS i.e. P+S is connected to N+body by a metal interconnect and N+body makes an ohmic contact to N-Well. This ensures that Source-Body bias is zero which leads to zero BODY EFFECT hence threshold voltage of the MOS is held constant.

The equivalent circuit model is given in Figure 7.3.16.3.

As seen in the Figure, the positive voltage bus acts as the Anode of the parasitic SCR and the GND bus i.e. Vss bus acts as the Cathode of the parasitic SCR. *For discussion purpose we assume that Rsubstrate and Rwell are infinite*. Under normal condtions voltage between the two buses is 5V or less hence SCR doesnot fire and there is no latching.

7.3.16.1. Why does SCR fire above break-over voltage ?

By inspection of the Circuit Model of SCR in Figure 7.3.16.3 we get the following equations:

$$I_{E1} = I_{C1} + I_{B1} \text{ , } I_{E2} = I_{C2} + I_{B2} \text{ and } I_{C1} = I_{B2} \text{ ,}$$

$$I_{C2} = I_{B1} \qquad 7.3.16.1$$

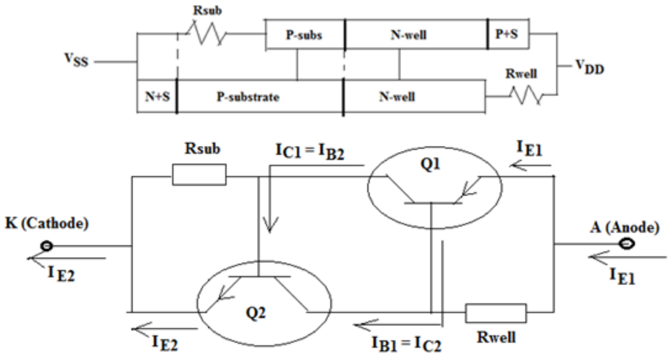


Figure 7.3.16.3. The Circuit Model of cross-coupled transistors Q1 and Q2 where Q1 is PNP transistor and Q2 is NPN transistor.

We further see that:

$$I_{C1} = \alpha_{F1} I_{E1} + I_{CBO1} \qquad 7.3.16.2$$

$$I_{C2} = \alpha_{F2} I_{E2} + I_{CBO2} \qquad 7.3.16.3$$

Where I_{CBO} = Collector-Base Junction reverse leakage current with Emitter open.

Adding the two above equations, we get:

$$I_{C1} + I_{C2} = \alpha_{F1} I_{E1} + I_{CBO1} + \alpha_{F2} I_{E2} + I_{CBO2} \qquad 7.3.16.4$$

By rearranging the terms we get:

$$I_{C1} + I_{C2} = \alpha_{F1} I_{E1} + \alpha_{F2} I_{E2} + I_{CBO1} + I_{CBO2} \qquad 7.3.16.5$$

We know that $I_{E1} = I_{E2} = I_E$ by Kirchoff's Current Law;

Also $I_{C1} + I_{C2} = I_{E2} = I_E$;

Therefore rearranging the terms we get:

$$I_E = \frac{I_{CBO1} + I_{CBO2}}{[1 - (\alpha_{F1} + \alpha_{F2})]} \qquad 7.3.16.6$$

A four layer device behaves like two transistors cross coupled as shown in the lower part of Figure 7.3.16.3. This four layer device behaves like Silicon Controlled Rectifier which has the symbol and I-V characteristics as shown below.

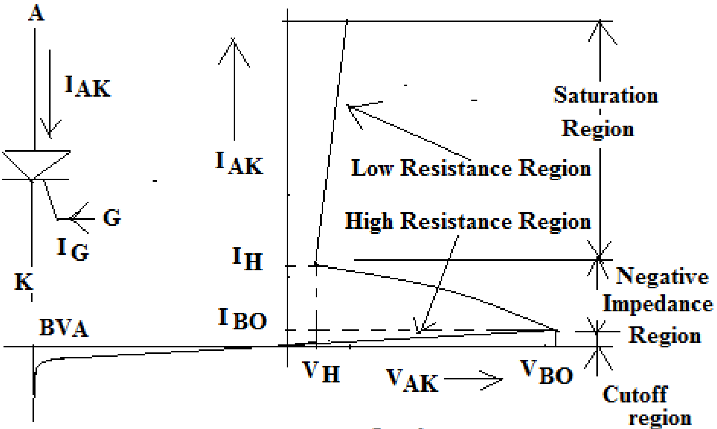


Figure 7.3.16. 4 The Symbol of SCR and Latch-up current versus voltage. Transition from high to low resistance when V_{AK} exceeds V_{BO} (break-over voltage) or when triggered by the gate current.

Forward Current Transfer Ratio of BJT is $\alpha_F = \gamma \times \beta^* \times M$

Where γ = injection efficiency. It falls at low currents due to recombination in depletion layer and it also falls at high current due to conductivity modulation. It is at the maximum value at moderate currents.

Also $\beta^* =$ base transport factor. This improves as Base Width becomes narrower which it will at higher voltages of V_{CE}(forward active mode).

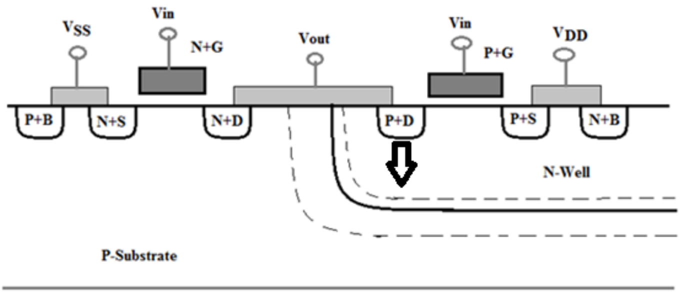
M= Avalanche Multiplication Factor which does not come into picture at low voltages.

Initially α_F is very small due to very low currents flowing through it but as V_{AK} (voltage between anode and cathode) increases, due to shrinking of Base width both alphas improve.

At $\alpha_{F1} + \alpha_{F2} = 1$, the current increases in a runaway fashion only limited by the external resistance. This is the firing point of the parasitic SCR embedded in the CMOS structure and can do irreversible damage to the IC.

7.3.16.2. How does the runaway process of SCR current set in?

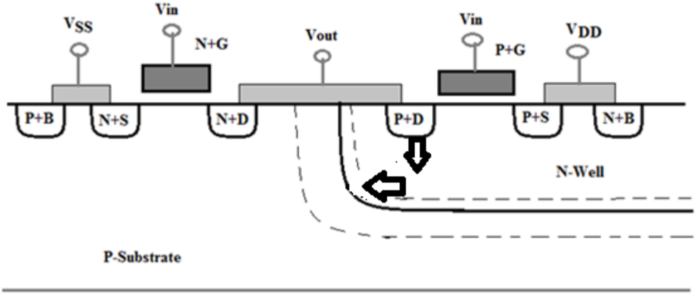
Consider the following inverter structure:



Step 1- PMOS P+ drain junction becomes forward bias during transient.
- holes are injected into N-Well as shown.

Figure 7.3.16.5. Triggering of CMOS Latch -up step 1.

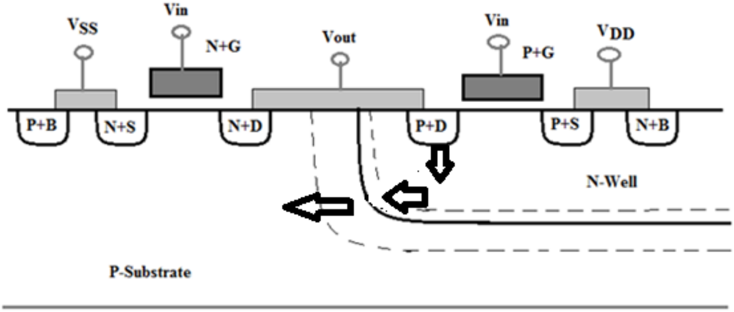
Consider the following inverter structure:



Step 2 - minority holes diffuse in N-Well. Some reach the N-Well & P-Subs depletion region.

Figure 7.3.16.5. Triggering of CMOS Latch -up Step 2

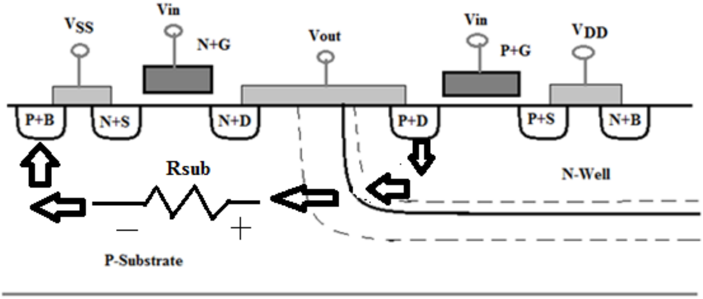
Consider the following inverter structure:



Step 3- built-in potential gradient in the depletion region is always a down-hill for minority carriers and up-hill for majority carriers. Hence built-in electric field sweeps the holes into P-substrate.

Figure 7.3.16.5. Triggering of CMOS Latch -up Step 3

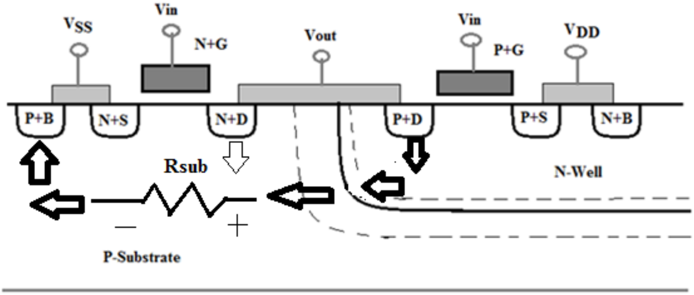
Consider the following inverter structure:



Step 4. These holes diffuse across P-subs and must leave through P+ body contact which in this case is P-substrate contact.
-------hole current through P-substrate causes voltage drop in substrate.

Figure 7.3.16.5. Triggering of CMOS Latch -up Step 4

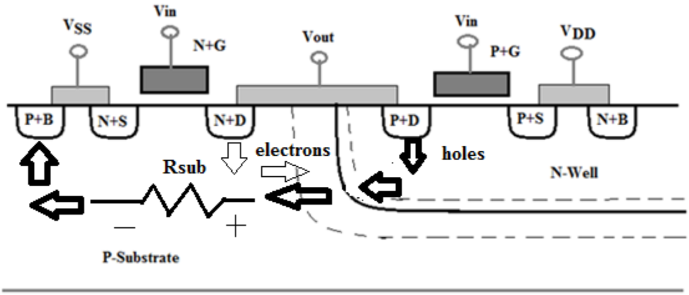
Consider the following inverter structure:



Step 5 - voltage drop across Rsub can forward bias NMOS drain -substrate junction.
---------- this forward bias causes electrons to be injected into substrate.

Figure 7.3.16.5. Triggering of CMOS Latch -up Step 5

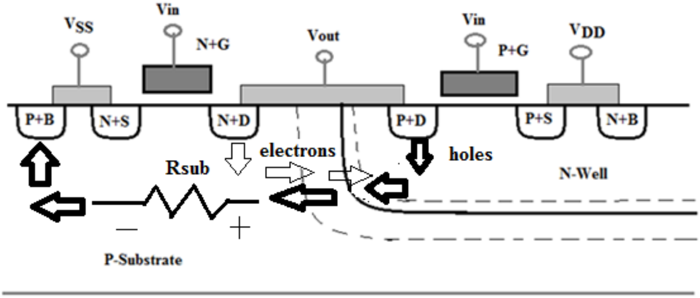
Consider the following inverter structure:



Step 6 - some of the electrons diffuse to depletion region edge.

Figure 7.3.16.5. Triggering of CMOS Latch -up Step 6

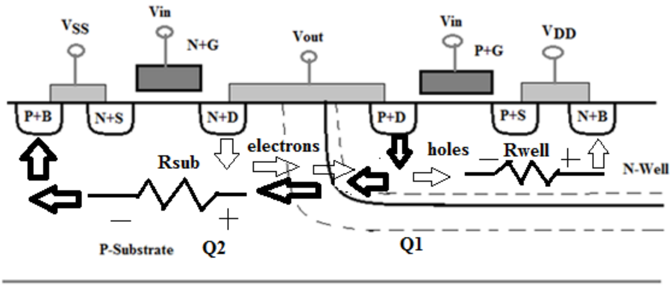
Consider the following inverter structure:



Step 7 - built-in field sweeps electrons across the depletiion region into N-Well.

Figure 7.3.16.5. Triggering of CMOS Latch -up Step 7

Consider the following inverter structure:



Step 8 - electrons must leave through N+ body (V DD contact).
---------- electron current through Rwell causes a voltage drop in N-Well.

Figure 7.3.16.5. Triggering of CMOS Latch -up Step 8

As we see in Figure 7.3.16.5. Step 8 a positive feed back loop or regenerative loop is set up.

A transient at P+Drain of PMOS sets up a hole current through Rsub which triggers N+drain of NMOS to set up an electron current through Rwell.

The electron current through Rwell further enhances hole current through Rsub and hole current through Rsub further increments the electron current through Rwell. This leads to a runaway condition which results in a large current flow from VDD bus to Vss bus and subsequently damages the CMOS structure.

7.3.16.3. The remedies of Latching Problem in N-Well process/P-Well process.

Design approach.(internal implants and larger spacing)

As we see in Figure 7.3.16.5. Step 8, we have two

BJTs Q1 and Q2.

Q1 is comprised of P+(drainof PMOS)-Nwell-Psubstrate and Q2 is comprised of N+(drain of NMOS)-Psubstrate-Nwell.

Q1 current gain can be reduced by introducing a buried layer in N-Well. This reduces the injection efficiency.

Q2 current gain can be reduced by increasing the spacing between N-Well and N+ Source/Drain. Increase in the spacing leads to reduced Base Transport Factor.

Fabrication approach.

As we saw in our discussion, Rsub and Rwell play a crucial role in setting up the runaway process. If by chosing higher doping level we reduce Rsub and Rwell then also runaway process will be prevented

Guard rings around N-Well with frequent contacts to the rings reduces the parasitic resistances. This also prevents the latch-up. The CMOS layout plan view is shown in Figure 7.3.16.6 and in Figure 7.3.16.7. without and with guard ring respectively.

7.3.16.4. Prevention of Latching in Twin-Tub Process.

In Twin-Tub process by the formation of ‘Refilled

Trench Isolation' between Twin Tub.This isolation prevents formation of parasitic BJT as shown in Figure 7.3.16.8.

The trench is formed in Silicon Substrate by anisotropic reactive sputter etching. Oxide layer is grown on the side-walls and bottom of the trench. This oxide lined trench is filled by Poly-Si or SiO_2 . This kind of structure is shown in Figure 7.3.16.8.

The trench isolation disrupts the parasitic BJT hence parasitic SCR is prevented structurally as shown in Figure 7.3.16.8

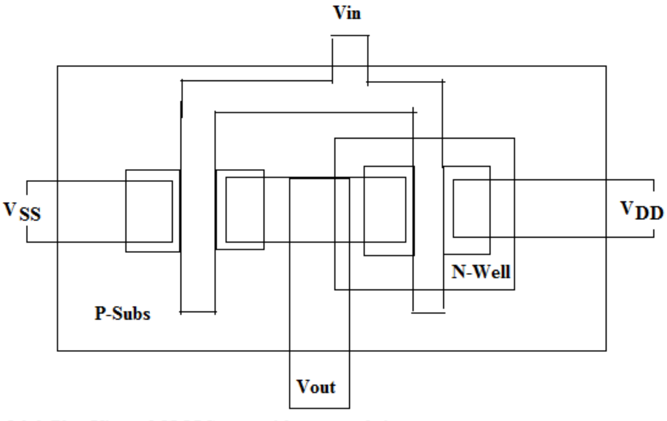


Figure 7.3.16.6. Plan View of CMOS gate without guard ring.

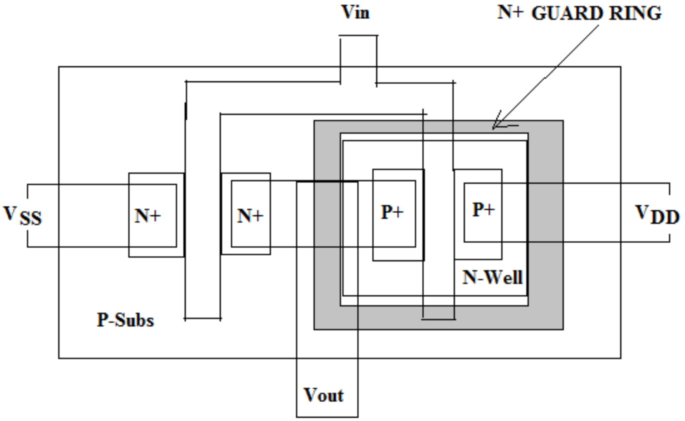


Figure 7.3.16.7. Plan View of CMOS with guard ring surrounding N-well

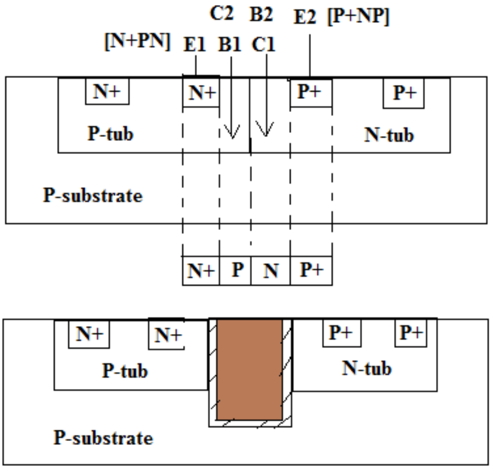


Figure 7.3.16.8.. Upper Figure is the cross sectional view of CMOS gate without trench isolation. Lower Figure is the cross sectional view of CMOS Gate with Refilled Tranch Isolation.

7.3.16.5. BiCMOS Latch-up susceptibility.

BiCMOS is less susceptible to latch-up problems because it has less substrate body resistance (Rsub) and less n-well body resistance (Rwell). Hence it requires larger currents through Rsub and Rwell to trigger latch-up process.

A typical BiCMOS in N-well process is shown in Figure 7.3.16.9

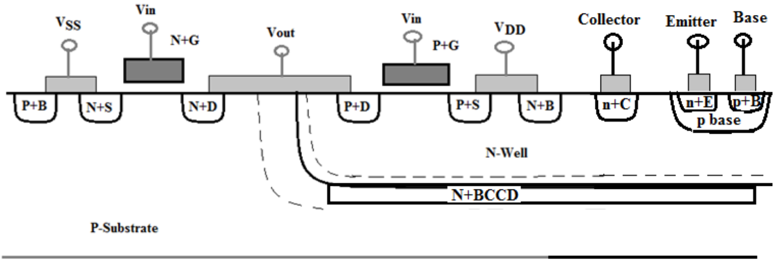


Figure 7.3.16.9 The cross-sectional view of N-Well BiCMOS

As can be seen from Figure 7.3.16.9 there is parasitic vertical PNP transistor constituted of

P base, N-Well and Psubstrate. This is a part of the N-Well Latch-up circuit. Due to the presence of N +BCCD(buried N+ subcollector), the life-time of minority carriers in the base of the parasitic vertical PNP is reduced. This reduces beta of the vertical PNP transistor since beta = life-time of minority cariers/ transit time. Hence latch-up susceptibility is reduced.

SSPD_Chapter 7_part 4_CMOS lay-out design rules.

SSPD_Chapter 7_Part 4 gives the importance of choosing the design rules for Minimum width, Minimum spacing, Surround and Extension.

SSPD_Chapter 7_part 4_CMOS lay-out design rules.

7.4.1. Introduction to VLSI Design.

Digital VLSI systems are mainly implemented by CMOS family of Logic in Silicon Wafer. Only for very high speed niche applications we go for ECL family of Logic. In this Chapter 7, we are mainly concerned with CMOS VLSI circuits. Figure 7.4.1 gives the flow chart of design and implementation of a CMOS Digital Systems.

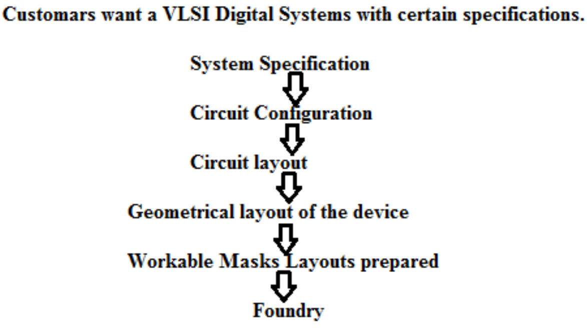

Customars want a VLSI Digital Systems with certain specifications.
System Specification
Circuit Configuration
Circuit layout
Geometrical layout of the device
Workable Masks Layouts prepared
Foundry

Figure 7.4.1. Flow Chart for design and implementation of VLSI Digital Systems.

7.4.2. Layout Design Rules.

As seen in Figure 7.4.1. from the designers layout the masks will be prepared and these masks will be translated into IC circuits in the foundry. These masks will be prepared within certain tolerances and with certain minimum resolutions. Mask preparation is the crucial step in IC technology and inorder that ICs are fabricated with high yield rate the tolerances and rsolutions must be clearly defined and hence layout design rules have been given.

Minimum wire widths are specified to avoid breaks. Minimum wire spacings are specified to avoid shorts. Overlaps are specified to ensure that overlap is complete. The design rules are expressed in terms of microns. The design rules are required so that a certain minimum resolution/tolerance of masks are permissible.

The Generation of the Fabrication Process technology is specified by the minimum feature size. Minimum feature size refers to channel length. As the Generation of the Fabrication Process advances so does the packing density and level of integration advances as shown in Table 7.4.1.

Table 7.4.1. Different Generations of Technology and their respective Packing Densities in Microprocessor Chips.

Year	Chip	L	Transistors per cm square
1971	4004	10µm	2.3K
1974	8080	6µm	6.0K
1976	8088	3µm	29K
1982	80286	1.5µm	134K
1985	80386	1.5µm	275K
1989	80486	0.8µm	1.2M
1993	Pentium	0.8µm	3.1M
1995	Pentium Pro	0.6µm	15.5M
1999	Mobile PII	0.25µm	27.4M
2000	Pentium 4	0.18µm	42M
2002	Pentium 4(N)	0.13µm	55M

There are different design rules such as: Mead-Conway Design Rules, CUMF n-MOS Design Rules and Lambda based Design Rules. In this lecture we will concentrate on Lambda Based Design Rules.

1.Design rules are a set of geometrical specifications that dictate the design of the layout masks.

2.A design rule set provides numerical values:

- For minimum dimensions
- For minimum line spacings

3.Design rules must be followed to insure functional structures on the fabricated chip.

4.Design rules change with technological advances (www.mosis.org).

Design rules are classified as :Minimum width, Minimum spacing, Surround and Extension.

The following physical limitations compel us to specify the design rules:

1. Line width limitation of an imaging system:
    - The reticle shadow projected on the photoresist does not have sharp edges due to optical diffraction.

3. Etching process problem:
    - Undercutting of the resist due to lateral etching decreases the resolution.

Limits on N+ spacing:

If depletion regions of adjacent pn junctions touch, then

- The current blocking characteristics are altered and
- Current can flow between the two.

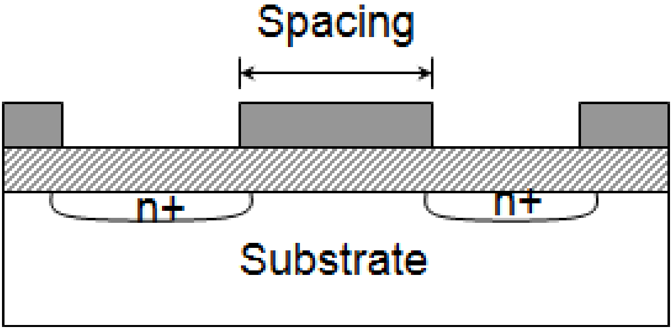


Figure 7.4.2. The depletion region surrounding N+ regions can merge hence a definite spacing has to be kept.

Electrical capacitive coupling can occur between closely spaced conducting lines. This leads to a problem called crosstalk

- A portion of the electrical energy is coupled to another causing noise.
- This is a major problem in high-density design.

Hence closely spaced metal lines are kept well separated.

Certain electrical rules are followed such as the allowed width of a metal interconnect line:

- To avoid electromigration effects
- The design rule set will stipulate the maximum current flow level permitted

3D Perspective

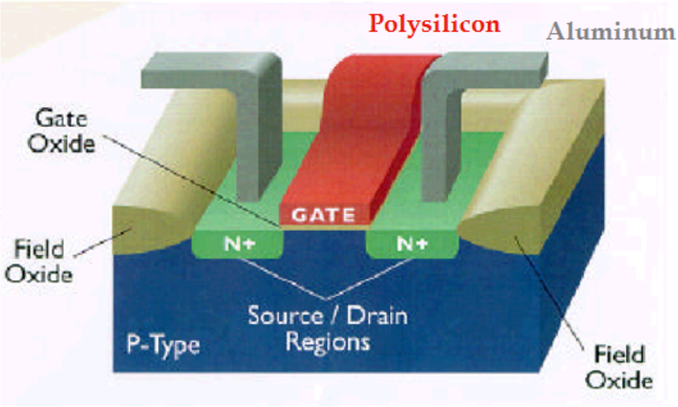


Figure 7.4.3. Three Dimensional Perspective View of NMOS fabricated in P-substrate.

Why do we use design rules?

The design rules provide the interface between designer and process engineer.

The design rules also provide the guidelines for constructing process masks.

Minimum length or width of a feature on a layer is 2lambda where lambda is half channel length. This is to allow for contraction of the feature.

Minimum separation of features on a layer is 2lambda.This is to ensure adequate continuity of the intervening materials.

SSPD_Chapter 7_part 4_CMOS lay-out design rules_continued2.

SSPD_Chapter 7_Part 4_continued2 gives the design rules for the minimum width for the metal interconnect,poly-silicon line and diffusion line.It gives the design rule for the spacing between metal-metal,poly-poly , diffusion-diffusion and between diffusion-poly.

SSPD_Chapter 7_part 4_CMOS lay-out design rules_continued2.

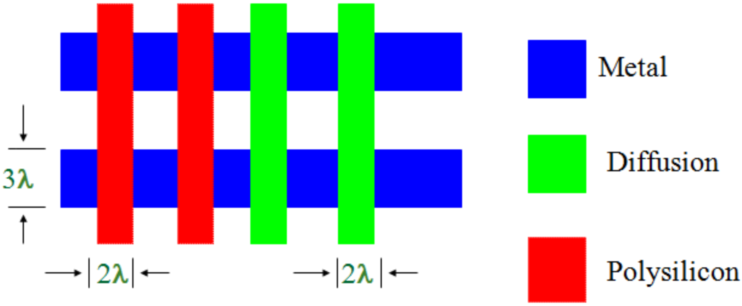
7.4.3.Lambda Based Design Rules.

Every Foundry follows a standard. The foundry allows the designers to submit designs using a state-of the-art process. They permit widths, spacings and distances on the layout to be written in the form mλ where λ = half channel length.

Design Rules

Minimum width of PolySi and diffusion line 2λ

Minimum width of Metal line 3λ as metal lines run over a more uneven surface than other conducting layers to ensure their continuity

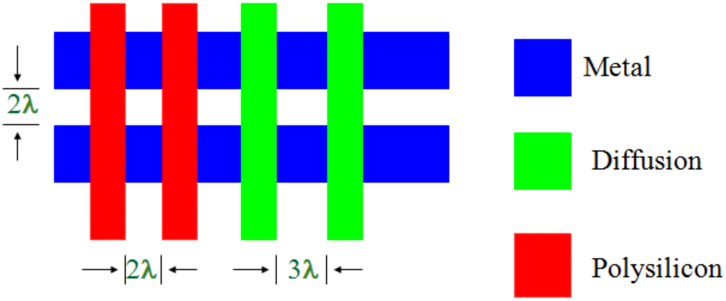


Design Rules

PolySi – PolySi space 2λ

Metal - Metal space 2λ

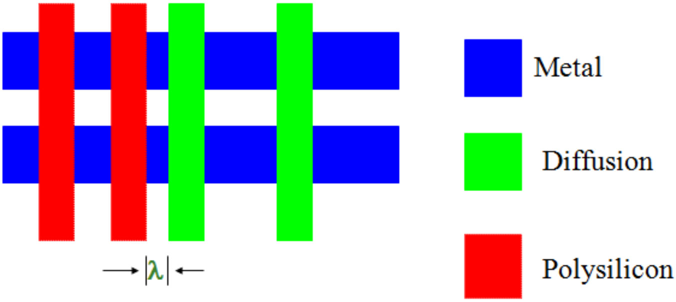
Diffusion – Diffusion 3λ **To avoid the possibility of their associated regions overlapping and conducting current**



Design Rules

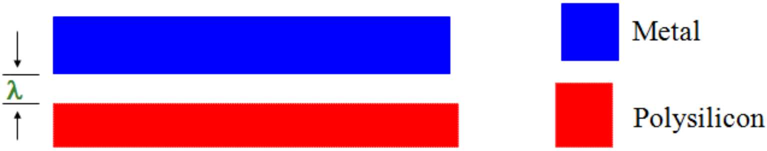
Diffusion – PolySi λ To prevent the lines overlapping to form unwanted capacitor

Metal lines can pass over both diffusion and polySi without electrical effect. Where no separation is specified, metal lines can overlap or cross



Metal Vs PolySi/Diffusion

- Metal lines can pass over both diffusion and polySi without electrical effect
- It is recommended practice to leave λ between a metal edge and a polySi or diffusion line to which it is not electrically connected



Review:

- **poly-poly spacing 2λ**
- **diff-diff spacing 3λ**
  (depletion regions tend to spread outward)
- **metal-metal spacing 2λ**
- **diff-poly spacing λ**

Note

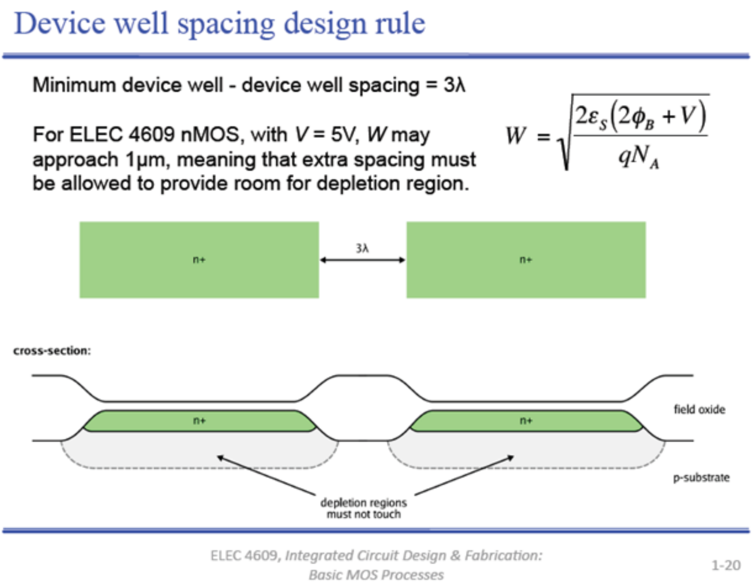
Two Features on different mask layers can be misaligned by a maximum of 2λ on the wafer.

- If the overlap of these two different mask layers can be catastrophic to the design, they must be separated by at least 2λ
- If the overlap is just undesirable, they must be separated by at least λ

SSPD_Chapter 7_part 4_CMOS lay-out design rules_concluded.
SSPD_Chapter 7_Part 4_CMOS Layout conclusion gives the design rules for contact enclosures, device well to device well spacing and the complete N-Well CMOS layout with the PMOS:NMOS =(W/L)p:(W/L)n = 2:1 for best noise margin.

SSPD_Chapter 7_Part 4_CMOS design rules_continued3

7.4.3.1. Device Well-Device Well spacing should be 3Lambda.



Device well spacing design rule

Minimum device well - device well spacing = 3λ

For ELEC 4609 nMOS, with V = 5V, W may approach 1µm, meaning that extra spacing must be allowed to provide room for depletion region.

$$W = \sqrt{\frac{2\varepsilon_S(2\phi_B + V)}{qN_A}}$$

ELEC 4609, *Integrated Circuit Design & Fabrication: Basic MOS Processes*

1-20

7.4.4. Metal Overlay on Diffusion Contact/Poly Si Contact/Metal 1,Metal 2 Via.

In Figure 7.4.4.1 we give the design rules for Contact Size and Contact Spacing with respect to Diffusion Well/PolySi Edge /Channel, minimum poly-spacing with respect to adjacent diffusion well, minimum gate overlap of Diffusion –Well and minimum Source/Drain width according to CUMF NMOS Design Rules. Contact area is marked black.

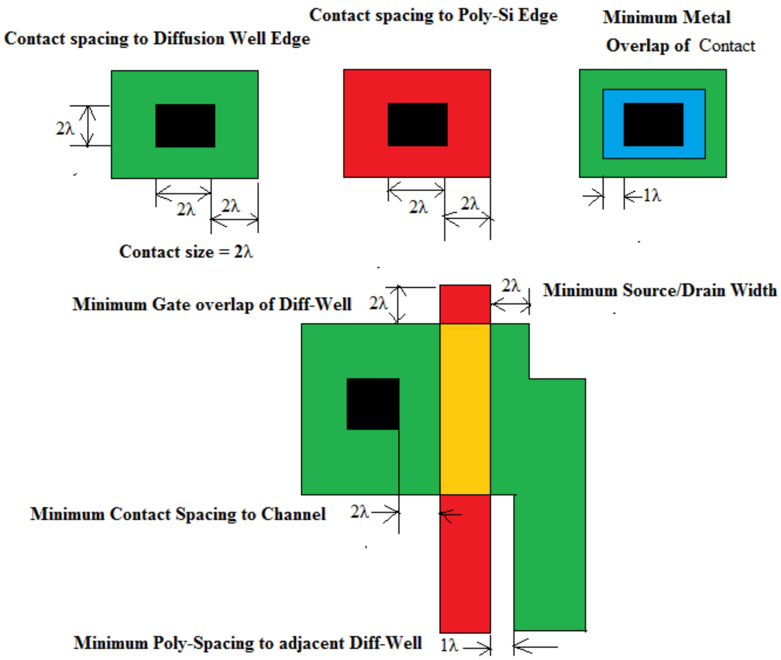

Contact spacing to Poly-Si Edge
Minimum Metal
Overlap of Contact
Contact spacing to Diffusion Well Edge
2λ
2λ
2λ
2λ
2λ
1λ
Contact size = 2λ
Minimum Gate overlap of Diff-Well
2λ
2λ
Minimum Source/Drain Width
Minimum Contact Spacing to Channel
2λ
Minimum Poly-Spacing to adjacent Diff-Well
1λ

Figure 7.4.4.1. The design rules for Contact Size and Contact Spacing with respect to Diffusion Well/PolySi Edge /Channel, minimum poly-spacing with respect to adjacent diffusion well, minimum gate overlap of Diffusion –Well and

minimum Source/Drain width according to CUMF NMOS Design Rules. Contact area is marked black.

In Figure 7.4.4.2 we give the definition of Via and how through Vias different level of metallization are interconnected. Minimum distance between two vias is 3Lambda.

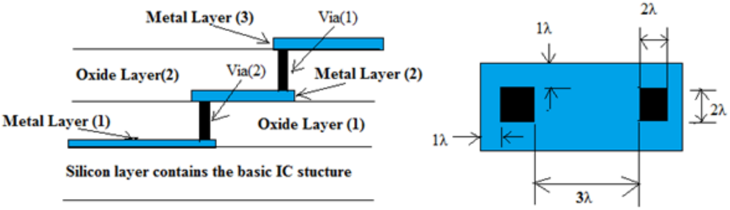


Figure 7.4.4.2. Definition of Vias and spacing between vias.

7.4.5. Rules for CMOS Layout.

CMOS can be fabricated by the following three methods:

1. N-Well in P-Substrate;
2. P-Well in N-Substrate;
3. Twin Well in P-Substrate.

We will choose N-Well in P-Substrate. In this case we have the following three requirements:

1. Definition of n-well area

2. Threshold implant of two types of transistor

3. Definition of source and drains regions for the NMOS and PMOS.

To ensure the separation of the PMOS and NMOS devices, n-well supporting PMOS is 6lambda away from the active area of NMOS transistor.

This is to prevent overlap of the associated regions as shown in Figure 7.4.5.1.

Simplest MOS processes require only 4 masks. It can be p-channel or n-channel FETs, not both.

In CMOS we have n and p transistors transistors simultaneously and therefore CMOS is a major improvement over NMOS or PMOS technology.

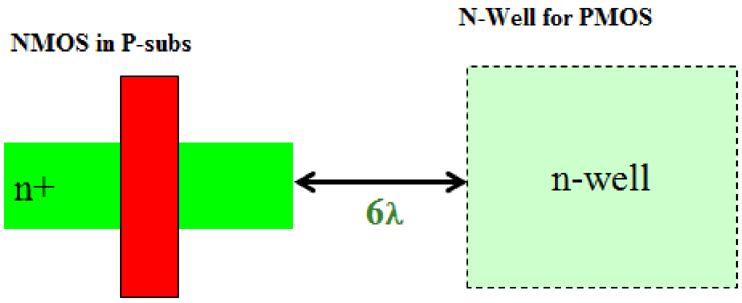


Figure 7.4.5.1 Seperation of NMOS and PMOS devices by 6lambda.

7.4.5.1. Complete CMOS Layout.

We have seen that for maximum noise margin, CMOS should have symmetrical operation around

$V_{in} = V_{out} = 0.5\ V_{DD}$. This requires that special transconductance parameter $\beta_N = \beta_P$. Since mobility of majority carriers in PMOS is half that of majority carriers in NMOS hence $(W/L)_P = 2\times\ (W/L)_N$. Based on this inference the CMOS layout is given in Figure 7.4.5.2.

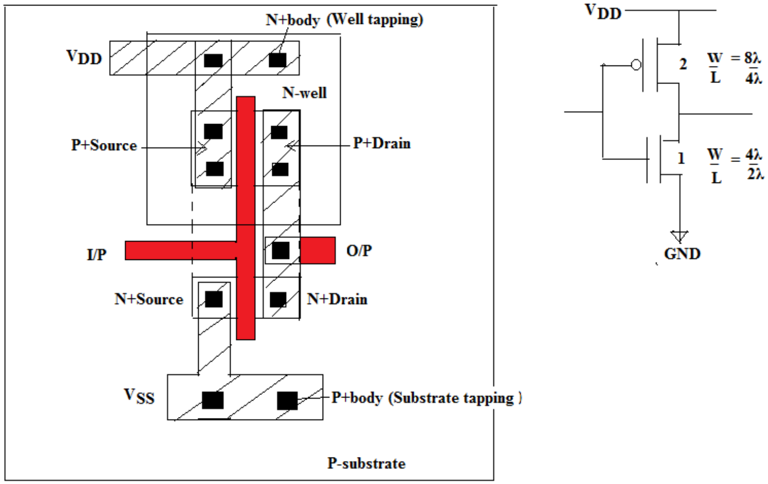


Figure 7.4.5.2. N-Well CMOS Layout and its geometry for optimum performance

SSPD_Chapter 7_Part 5_Stick Diagram of Logic Gates. SSPD_Chaptr 7_Part 5 introduces the concept of stick diagram in context of VLSI design.

SSPD_Chapter 7_Part 5_Stick Diagram of Logic Gates.

7.5 What is a Stick Diagram of a logic gate.

VLSI circuit is a 3-D set of patterned layers. Stick Diagram provides a top view of the patterns. Colours allow us to trace signal flow paths through conduction layers. It is a schematic of a circuit at physical design level.

Stick Diagram is a cartoon of the actual circuit. It is a circuit schematic with topology information but it has no real device size. It is not to scale and it uses color pencils to show the different layers of the device. Complicated wiring of gates and cells is easier to visualize with stick diagram. Planning and Physical Design using Stick Diagrams before going to CAD can save energy and time.

In Figure 7.5.1. the perspective view of (E)NMOS fabricated in P-Substrate is shown. At the bottom of the Figure 7.5.1. the top view of (E)NMOS fabricated in P-Substrate is given. The top-view contains the layout of the different layers of (E)NMOS:

Lowest layer is P-Substrate given by brown enclosure.

Second layer is N+ diffusion shown by green enclosure.

Topmost layer is Poly-Si shown by red colour.

The symbol, the layout and the stick diagram of (E)NMOS is given in Figure 7.5.2.

In Figure 7.5.3. the layout and the stick diagram of (E)PMOS in N-Well is shown.

In Figure 7.5.4, the contact or VIA needed at two adjacent levels of metallization is shown.

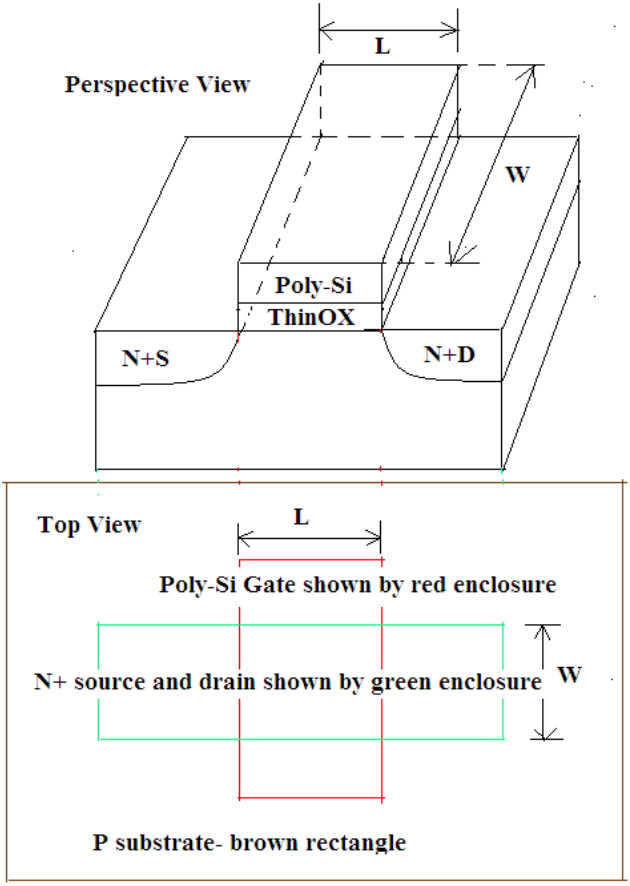

Perspective View
L
W
Poly-Si
ThinOX
N+S
N+D
Top View
L
Poly-Si Gate shown by red enclosure
N+ source and drain shown by green enclosure
W
P substrate- brown rectangle

Figure 7.5.1. Perspective view and top view of NMOS fabricated in P-substrate.

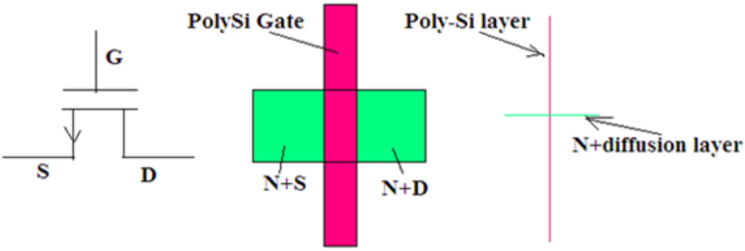


Figure 7.5.2 Symbol of NMOS, layout of NMOS and Stick-Diagram of NMOS

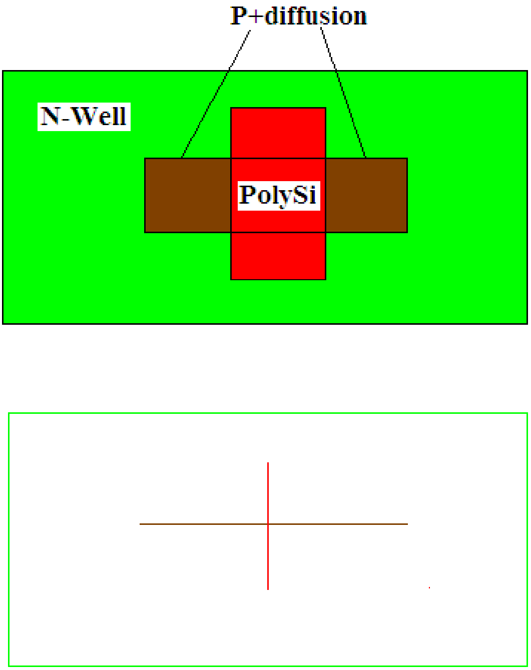


Figure 7.5.3. The layout and the stick diagram of PMOS in N-Well.

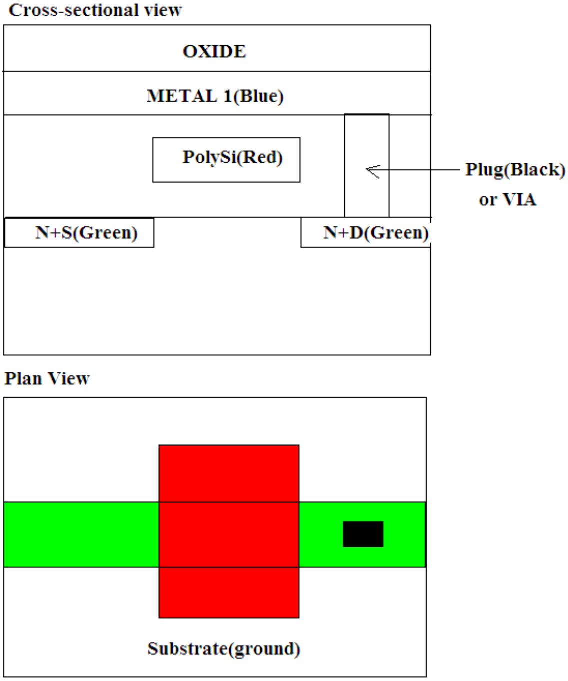


Figure 7.5.4. The cross-sectional view of the interconnection made between different layers of metallization through contact or VIA and the Plan View of the same with VIA marked black.

SSPD_Chapter 7_Part 5_Stick Diagrams of Logic Gates_continued 2

SPD_Chapter 7_Part 5 continues with simple examples of stick diagrams of (E)NMOS inverter with (D)NMOS pull up and using the same inverter for generating the complement of (A+BC).

SSPD_Chapter 7_Part 5_Stick Diagrams of Logic Gates_continued 2

7.5.1. NMOS inverter with (D)NMOS as pull-up transistor.

In Figure 7.5.1.1. we give the circuit diagram of (E)NMOS inverter with (D)NMOS as pull-up transistor.

In Figure 7.5.1.2. we give the circuit diagram of (E)NMOS inverter with (D)NMOS as pull-up transistor along with a Pass Transistor at the INPUT.

In Figure 7.5.1.3. we give the circuit diagram of Z = Complement of (CB + A) and its stick diagram.

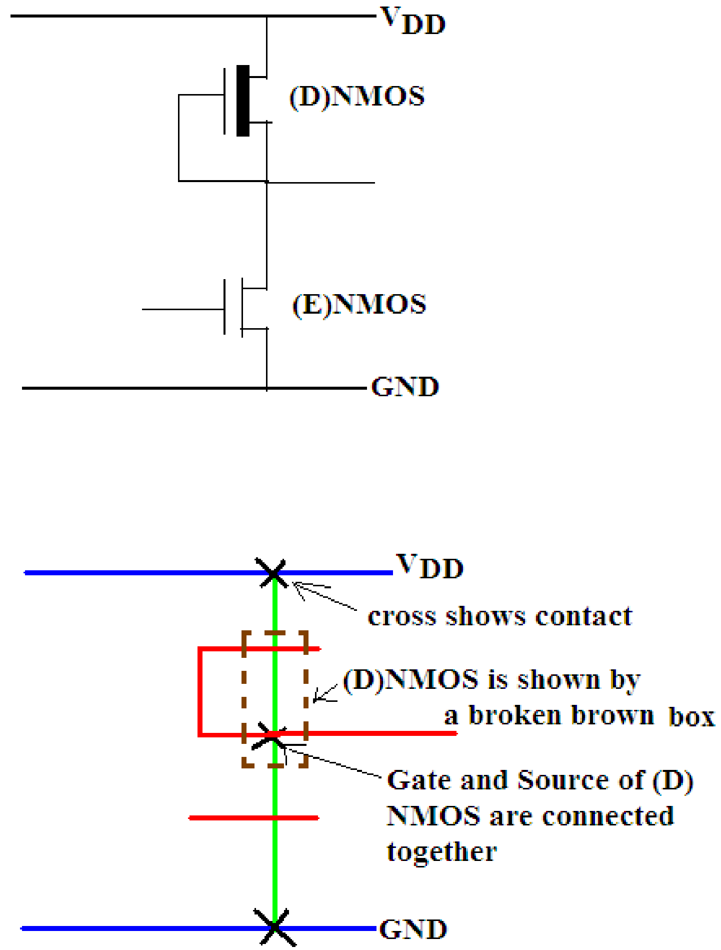
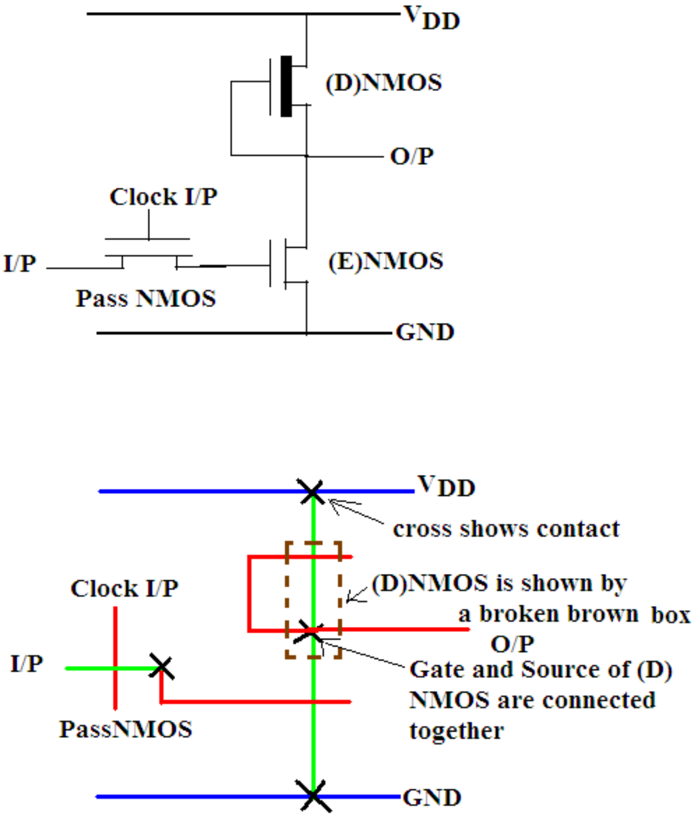

VDD
(D)NMOS
(E)NMOS
GND
VDD
cross shows contact
(D)NMOS is shown by a broken brown box
Gate and Source of (D) NMOS are connected together
GND

Figure 7.5.1.1. (E)NMOS Inverter with (D)NMOS as pull-up transistor.

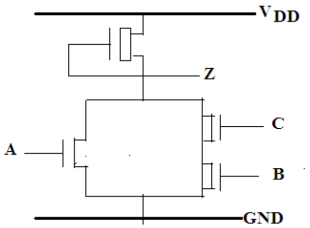


cross shows contact

(D)NMOS is shown by a broken brown box

Gate and Source of (D) NMOS are connected together

Figure 7.5.1.2. (E)NMOS inverter with (D)NMOS as pull-up transistor and with Pass Transistor at the input.



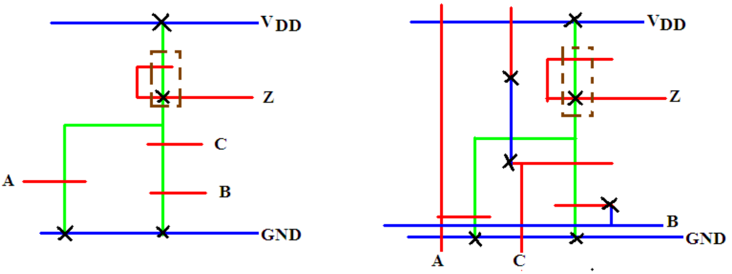


Figure 7.5.1.3. Two alternative forms of stick diagrams for Z=Complement(A+CB)

In all these three diagrams, metal layer is the topmost layer, Poly-Si is intermediate layer and diffusion is the lowest layer. So crossing doesnot imply contact. Only when contact is made at the intersection of different layers that contact is made between different layers.

Also (E)NMOS is simply red line crossing green line but (D)NMOS is red line crossing green line within a broken brown box. This implies that N implant has been made under the Gate of (D)NMOS so that the transistor is normally-on.

SSPD_Chapter 7_Part 5_Stick Diagram of CMOS Logic Gates_continued 3

SsPD_Chapter 7_Part 5_continued 3 gives the stick diagrams of CMOS inverter and two-input CMOS NAND Gate.

SSPD_Chapter 7_Part 5_Stick Diagram of CMOS Logic Gates_continued 3

7.5.2. Stick Diagrams of CMOS Logic Gates.

In Figure 7.5.2.1 gives three alternative layouts of CMOS Inverter expressed by its corresponding stick diagram.

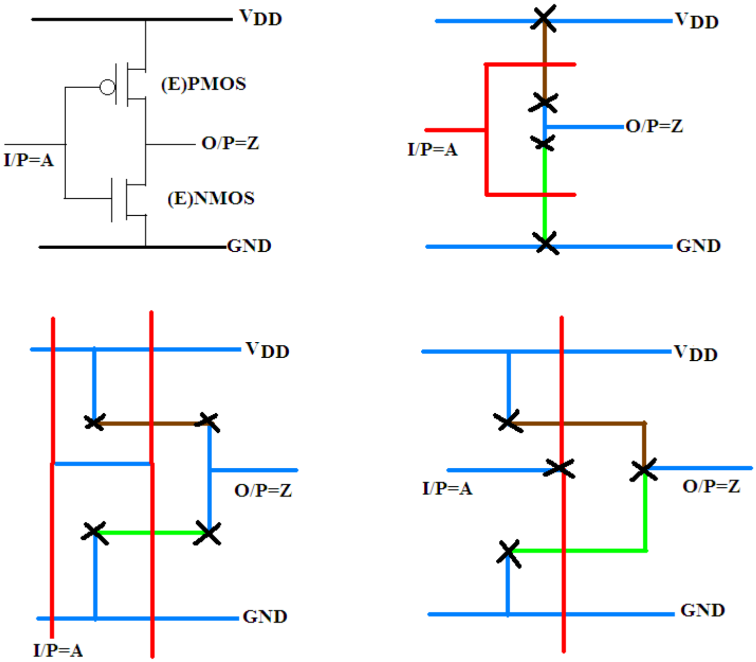


Figure 7.5.2.1. Three alternative layout of CMOS Inverter expressed by its corresponding stick diagrams.

Blue colour- metal1, Red – Poly-Si, Green – N+ diffusion, Brown – P+ duffusion, Black Cross is CONTACT.

7.5.3. CMOS Two-input NAND Gate

Figure 7.5.3.1 gives the circuit diagram and its alternative layout expressed in stick diagrams for a two-input NAND Gate.. Output Z = Complement(AB)

This means when A=1 and B=1 output is connected to GND bus by a low reesistance through Q1(ON) and Q2(ON) in series. When either of the two I/Ps is at ‘0’ low resistance path to GND is broken and a low reesistance path exists through Q3 or Q4 to 5V(Vdd) bus. Thus the condition for Output Z = Complement(AB) is satisfied.

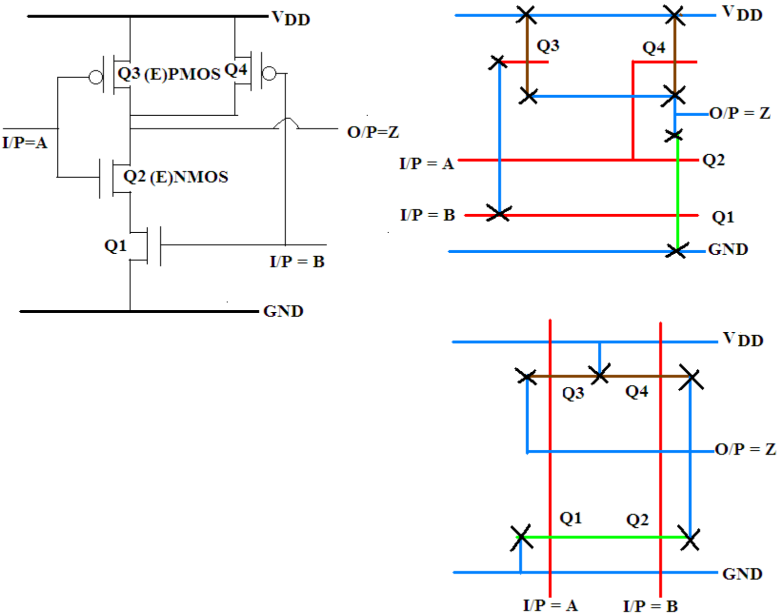


Figure 7.5.3.1 Two-Input CMOS NAND Gate.

In Figure 7.5.3.1, Q3 input is shorted to Q1 input and Q4 input is shorted to Q2. In the upper stick diagram, Q4 input and Q2 inpt are connected by the same Red Poly-Si line. But Q3 and Q1 input cannot be connected in the same way. If Q3 and Q1 inputs were connected by the same red line then it would short to the input of Q2 also which it should not. Therefore a metal line blue line is used to cross the red input of Q2. Metal blue line lies at an upper layer while red Poly-Si line lies at intermediate level therefore cross-over is permitted. Once the cross-over has taken place then the two terminals of blue metal line is connected to the two Poly-Si red inputs of Q3 and Q1.

Similarly 3-input NAND Gate can be designed. In case of 3-input NAND Gate we have Q1,Q2 and Q3 in series and Q4,Q5 and Q6 in parallel.

Q1 and Q4 are paired as CMOS, Q2 ans Q5 are paired as CMOS and Q3 and Q6 are paired as CMOS.

SSPD_Chapter 7_part 5_Stick Diagrams of CMOS Gates_concluded.

SSPD_Chapter 7_part 5_Stick Diagrams of CMOS Gates_concluded. is the concluding part of the stick diagram of CMOS Gates.It gives the layout of transmission gate,multiplexer and metal interconnect.

SSPD_Chapter 7_part 5_Stick Diagrams of CMOS Gates_concluded.

7.5.4. CMOS Transmission Gate.

Figure 7.5.4.1 gives the layout of transmission switch and its stick diagram.

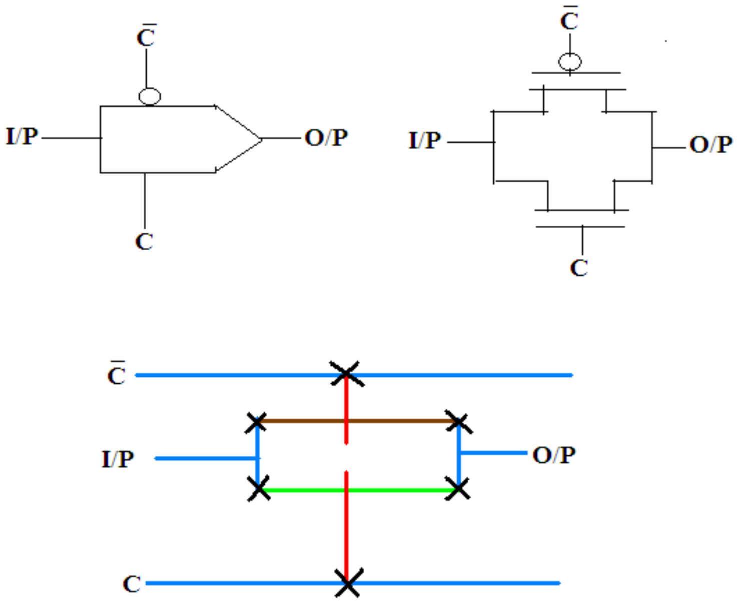


Figure 7.5.4.1. CMOS Transmission Gate circuit diagram and its layout expressed in stick diagram.

Blue- metal 1, Red – Poly-Si, Green – N+diffusion for (E)NMOS; Brown – P+diffusion.

Whenever control I/P ‘C’ = 1, ‘C complement’=0. Both the transistors turn ON and there is full transmission from I/P to O/P.

Whenever ‘C’= 0, ‘C complement’= 1 then both transistors turn ‘OFF’ and the transmission is stopped between I/P and O/P.

7.5.5. CMOS 2-input Multiplexer.

Figure 7.5.5.1 gives the Symbol and the layout of 2-input multiplexer expressed as stick disagram.

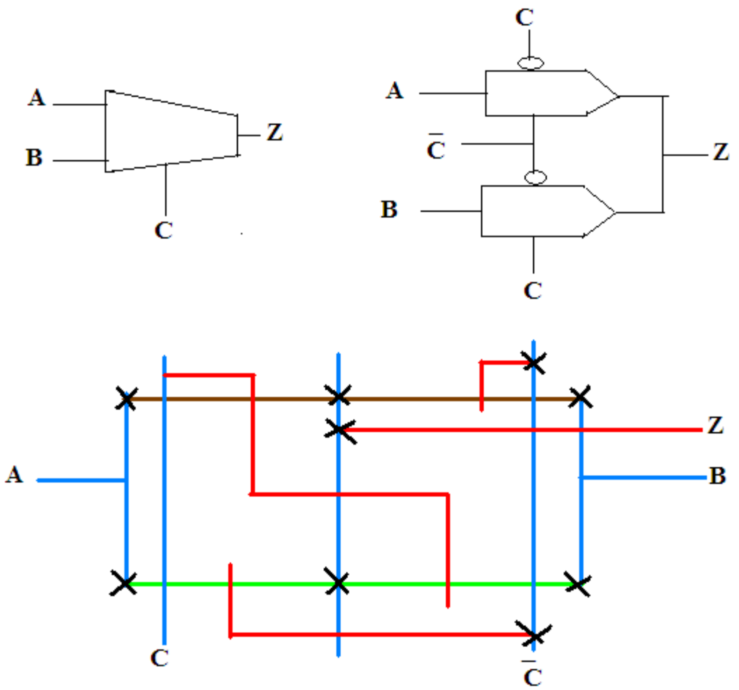


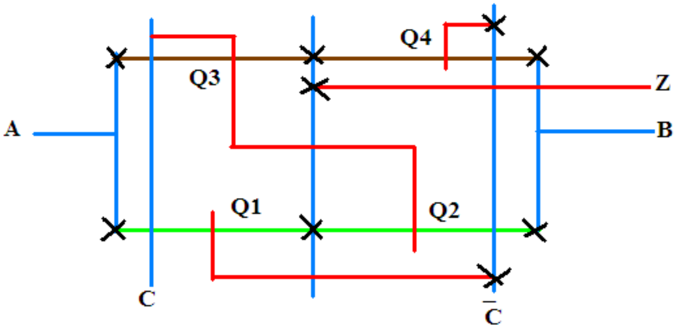
Figure 7.5.5.1.Two Input Multiplexer or 1-out-of-2 selector.

The circuit model of the layout is given in Figure 7.5.5.2.

By inspection of the circuit model we see that

when control input C=’1’ then Z = B and

when C= ‘0’ then Z = A.



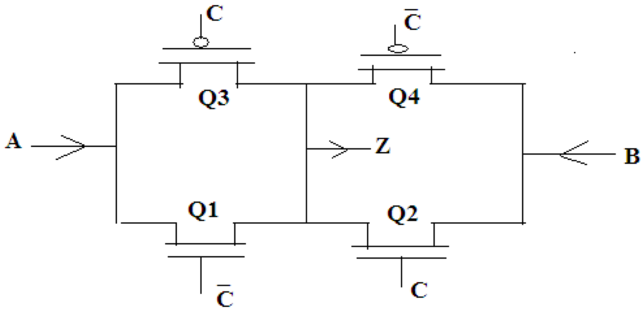


Figure 7.5.5.2. Equivalent Circuit Model from the layout of 2 input multiplexer.

- Metal Interconnect Layers.

Metal layers are electrically isolated from each other. Electrical contact between

adjacent conducting layers requires contact cuts and vias.

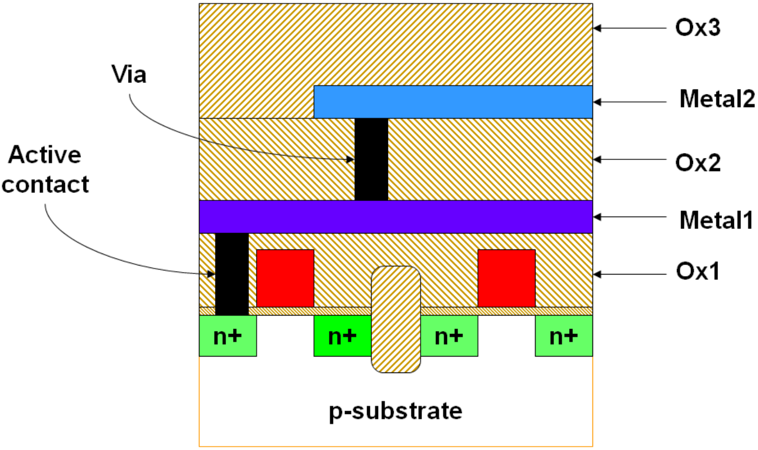


Figure 7.5.6.1. Cross-sectional view of two layers of metal and their interconnect.

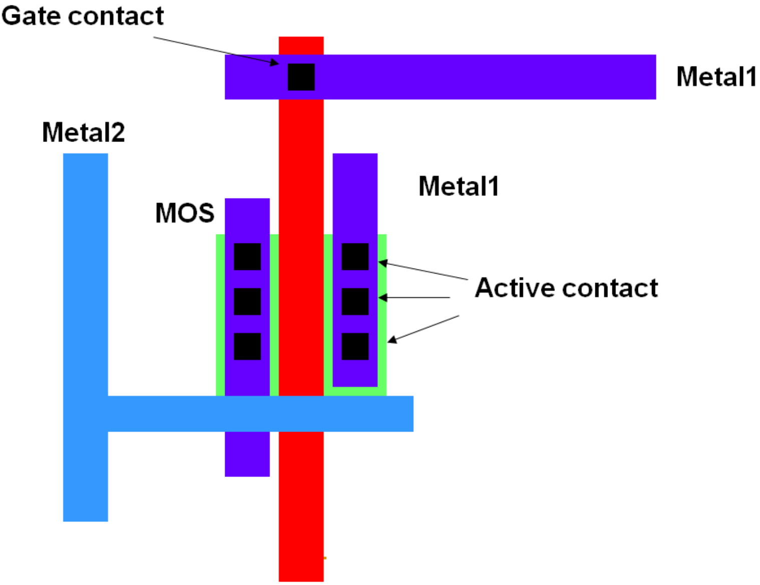


Figure 7.5.6.2. Plan View of intereconnect layout.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts. SSPD_chapter 7_Part 6 introduces the concept of Sheet Resistance and Capacitance per Unit Area. From the basic circuit concepts we can arrive at the absolute values of the circuit elements which help in modeling the circuit behaviour.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts.

7.6.1. Sheet Resistance.

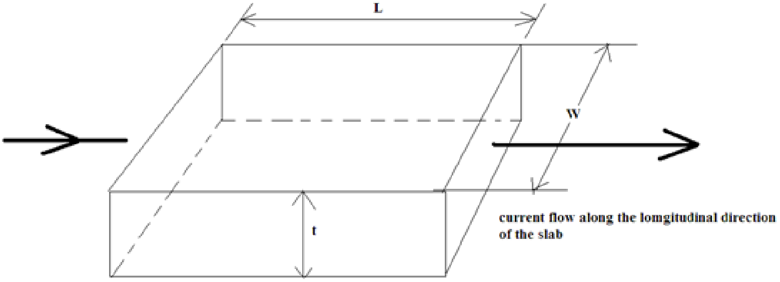


Figure 7.6.1. Resistance along the longitude of a rectangular slab.

Referring to Figure 7.6.1 we see that:

Resistance of the slab along the longitudinal axis:

$$R = \frac{\rho L}{A} = \frac{\rho \times L}{W \times t} = \frac{\rho}{t} \times \frac{L}{W} = R_{sheet} \times \frac{L}{W} \qquad 7.6.1.1$$

Therefore

$$R_{sheet} = \frac{\rho}{t} \quad (\frac{ohms}{square}) \qquad 7.6.1.2$$

Table 7.6.1 gives the typical values of the sheet resistance of the different layers of CMOS IC circuit.

Table 7.6.1.Typical values of the sheet resistance RS of MOS layers for technology generation 5µm, 2µm and 1.2µm. (Here 5µm, 2µm and 1.2µm imply the minimum feature size = channel length for the given technology . Consequently λ is 2.5µm, 1µm and 0.6µm for these three generations respectively).

Layer	Rs(Ω/ square)		
	5µm	2µm	1.2µm
Metal	0.03	0.04	0.04
Diffusion	10→50	20→45	20→45
Silicide	2→4		
Poly_Si	15→100	15→30	15→30
N channel	10^4	2×10^4	2×10^4
P channel	2.5×10^4	4.5×10^4	4.5×10^4

7.6.1.1. Resistance calculation of N-Channel in 5µm Generation of Technology.

If L:W = 2λ:2λ then $R_{channel} = R_{sheet} \times (L/W) = 1 \times 10^4 \Omega$..

If L:W = 8λ:2λ then $R_{channel} = R_{sheet} \times (L/W) = 1 \times 10^4 \Omega. \times 4 = 4 \times 10^4 \Omega.$.

With scaling, feature size has decreased and the pitch of metal interconnect has decreased. This has led to excessive RC time delay. To control this, metal interconnects are being replaced by Metal Silicides. The metal silicides or the refractory silicides are formed by depositing metal on poly-silicon and then sintering it. Though these have proved to be successful substitutes of metal interconnects the extra processing steps make it uneconomical.

With scaling the cross-sectional area of the metal interconnect decreases and current density increases. As this exceeds the limit of 1.5mA/mm, metal grains are pushed around and wire interconnects break. This is called migration. Silicides prevent electron migration.

With the latest generation of technology, silicides are being further replaced by copper interconnects.

7.6.2. Area Capacitance of Layers.

Parallel plate capacitance is given by the formula:

$$C = \frac{\varepsilon_0 \varepsilon_{ins} A}{D} \quad Farads \qquad 7.6.2.1$$

Where D = thickness of the dielectric between the plates;

A = cross-sectional area of the plate;

Absolute permittivity (permittivity) = ε_0 = 8.85×10^{-14} F/cm;

Relative permittivity of the dielectric in this case SiO_2 = 3.9;

In VLSI design we express Capacitance per unit area = (pF/μm^2). To obtain it in this form we have to the scaling factor as follows:

$$C_g\left(\frac{pF}{\mu m^2}\right) = \frac{\varepsilon_0 \varepsilon_{ins}}{D}\frac{F}{cm^2} \times \frac{10^{12} pF}{F} \times \frac{cm^2}{10^8 \mu m^2} =$$

$$\frac{35.416 \times 10^{-14}}{D} \times \frac{F}{cm^2} \times \frac{10^{12} pF}{F} \times \frac{cm^2}{10^8 \mu m^2};$$

$$D(cm) = \frac{35.416 \times 10^{-14}}{C_g \frac{pF}{\mu m^2}} \frac{F}{cm} = \frac{35.416 \times 10^{-14}}{C_g \frac{10^{-12} F}{10^{-8} cm^2}} \frac{F}{cm} = \frac{35.416 \times 10^{-14}}{C_g 10^{-4}} cm = \frac{35.416 \times 10^{-10}}{C_g} cm$$

But Gate-to-Channel capacitance, C_g (pF/μm^2) = 4×10^{-4}, 8×10^{-4}, 16×10^{-4} ;

Therefore D = 88.5nm, 44.27nm, 22.1nm

Typical values are given in Table 7.6.2.

Table 7.6.2. Typical area capacitance values for MOS cuircuits.

Capacitance	Value in pF × (10^{-4}/µm^2) (relative value in brackets)		
	5µm	2µm	1.2µm
Gate Oxide Thickness	88.5nm	44.27nm	22.1nm
Gate to Channel	4 (1.0)	8 (1.0)	16 (1.0)
Difusion(active)	1 (0.25)	1.75 (0.22)	3.75 (0.23)
Poly-Si to substrate	0.4 (0.1)	0.6 (0.075)	0.6 (0.038)
Metal 1 to substrate	.3 (0.075)	0. 33 (0.04)	0.33 (0.02)
Metal 2 to substrate	0.2 (0.05)	0.17 (0.02)	0.17 (0.01)
Metal 2 to metal 1	0.4 (0.1)	0.5 (0.06)	0.5 (0.03)
Metal 2 to poly-Si	.3 (0.075)	0. 3 (0.038)	0.3 (0.018)

7.6.2.1. Calculation of Standard Unit of Capacitance □C g

Generally while designing we donot work in terms of absolute values of capacitances. Instrtead we work in relative terms with repect to some standard value and the absolute value of that standard is recorded accurately.

In the present case the gate-to-channel capacitance of a MOS transistor with W = L = 2λ is calculated fot the different generations of technology and this is refered to as □C_g. Table 7.6.3 gives the values of □C_g for three generations of technology namely 5µm, 2µm and 1.2µm.

Table 7.6.3. Standard Unit of Capacitance for 5µm, 2µm and 1.2µm Generation.

	5µm	2µm	1.2µm
Gate-to-Channel □C_g	4 pF × 10^{-4}/µm^2 (5µm × 5µm) × 4 pF × 10^{-4}/µm^2 = 0.01 pF	8 pF × 10^{-4}/µm^2 (2µm × 2µm) × 8 pF × 10^{-4}/µm^2 = 0.0032 pF	16 pF × 10^{-4}/µm^2 (1.2µm × 1.2µm) × pF × 10^{-4}/µm^2 = 0.0023 pF

In Table 7.6.4 the oxide thickness is given for different Generations of Technology.

Table 7.6.4 Oxide Thickness scaling with channel length scaling.

L	5 µm	2 µm	1.2 µm	90nm	45nm	30nm
D	88nm	44nm	22nm	1.3nm = 5atomic layers		

As oxide becomes thinner, current leakage through the oxide layer by means of quantum tunneling and consequent heating becomes a serious problem. We need physically thick oxide to prevent the leakage and electrically thin to keep the threshold Voltage low. Rewriting Eq 7.6.2.1 we get:

$$C = \frac{\varepsilon_0 \varepsilon_{ins} A}{D} \quad Farads$$

By inspection of the above Equation we see that if relative permittivity is doubled in the numerator then thickness of the oxide layer,D, will also double thus making the Gate dielectric layer thick enough to prevent leakage aand leaving the threshold voltage unaffected.

Along with vertical scaling, lateral scaling has taken place. This has led to reduction in cross-sevtional area ‘A’. This decreases Capactance C and increases threshold voltage.. If threshold voltage has to be maintained around 0.5V then the thickness D will once again have to be decreased.

Therefore High K-metal gate was introduced. High K made the oxide layer thicker and metal gate prevented Fermi level pinning..

7.6.2.2. Metal 1 to Substraate Area Capacitance Calculation for 5μm Generation.

In Figure 7.6.2, the cross-sectional view and the plan view of a metal interconnect is shown.

The length by breadth of the metal track $=(L=20\lambda \times W=3\lambda)$ is given. Determine Metal 1 to Substrate area capacitance in terms of the standard capacitance.

From Table 7.6.1, for 5μm Generation of CMOS:

Metal 1 to substrate capacitance per unit area $= 0.075 \times \square C_g$

Therefore total area relative capacitance :

$$Total\ C_{m-sub} = \frac{20\lambda \times 3\lambda}{2\lambda \times 2\lambda} \times 0.075 \times \square C_g$$

Therefore Total area capacitance between metal 1 and substrate

$= 15 \times 0.075 \times \square C_g = 1.125 \square C_g$

The absolute value of the interplate capacitance

between metal 1 and substrate $= 1.125 \times 0.01\text{pF} = 0.0125\text{pF}$.

Similarly same area between poly-Si and substrate $= 15 \times 0.1 \times \square C_g = 1.5 \square C_g$;

Same area between n-type diffusion and substrate $= 15 \times 0.25 \times \square C_g = 3.75 \square C_g$;

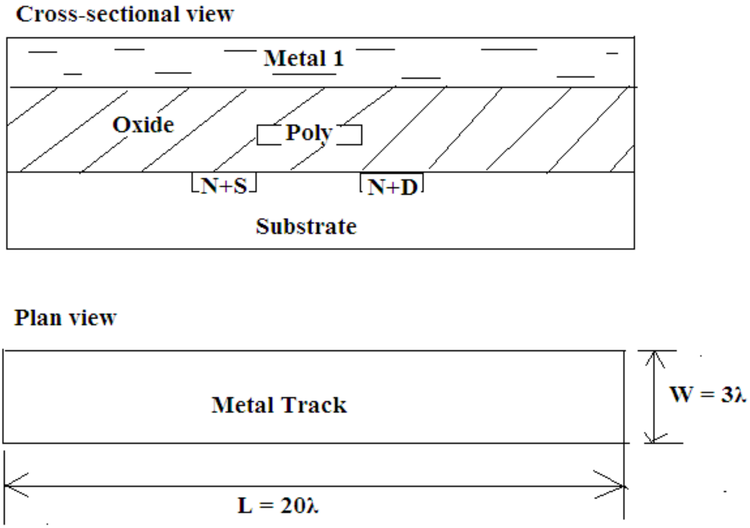


Figure 7.6.2.1 Calculation of area capacitance between metal layer 1 and substrate of the metal interconnect as shown.

7.6.2.3 Capacitance calculation in multi layer structures.

A CMOS IC with multi-layer metal interconnect is shown in Figure 7.6.2.2 and in Figure 7.6.2.3. plan view of metal interconnect making contact with Poly-Si extension from the gate is shown.

It should be noted that Poly-Si over the Gate lies over 88.5nm thin oxide layer has the nomenclature of Gate-to-Chanel and has the standard value of 1□C_g on the other hand

Poly-Si over the remaining portion runs over 5µm thick oxide layer and has the nomenclature Poly-Si-to-Substrate and has the value of 0.1□C_g.

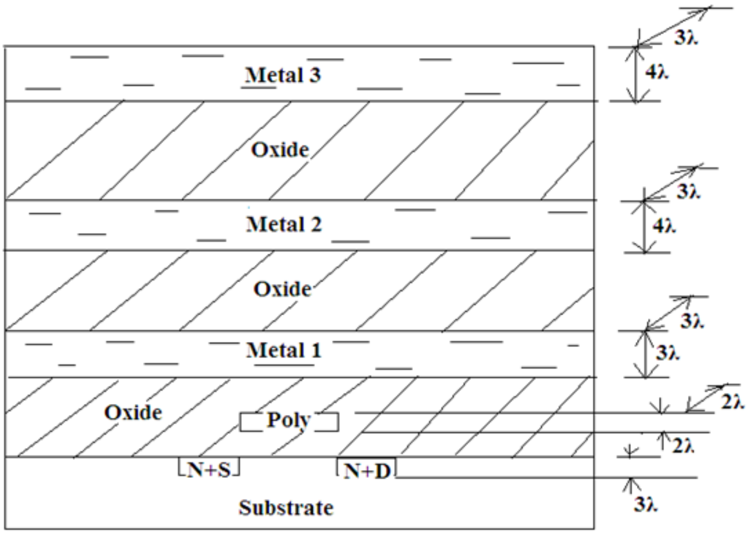


Figure 7.6.2. 2 Cross-sectional view of multi-layer structure.

thickness of M1=3λ and width of M1=3λ;
thickness of M2=4λ and width of M2=3λ;
thickness of M3=4λ and width of M3=3λ;

In Figure 7.6.2.3 we are required to determine Metal interconnect-to-Substrate capacitancve, Poly-Si-to-Substrate capacitance excluding Gate, Gate capacitance and total capacitance are to be calculated.

Relative area of metal interconnect with respect to standard capacitance $=(100\lambda\times 3\lambda)/(2\lambda\times 2\lambda) = 75$

Therefore Metal interconnect-to-Substrate Capacitance $= 75\times 0.075\square C_g = 5.625\square C_g$;

$C_m = 5.625\square C_g$;

Poly-Si-to-Substrate Capacitance(excluding Gate area)

$= [(\text{overlap area} + \text{extension area of Poly-Si excluding Gate})/(2\lambda \times 2\lambda)] \times 0.1$□$C_g$

$= [(4\lambda \times 4\lambda + 3\lambda \times 3\lambda)/ 2\lambda \times 2\lambda)] \times 0.1$□$C_g = 0.55$□$C_g$;

$C_{poly} = 0.55$□C_g;

$C_{gate} = 1$□C_g;

Therefore $C_{total} = C_m + C_{poly} + C_{gate} = (5.625 + 0.55 + 1)$ □$C_g = 7.175$□C_g;

The absolute value is $C_{total} = 7.157 \times 0.01\text{pF} = 0.0717\text{pF}$ for 5μm Generation.

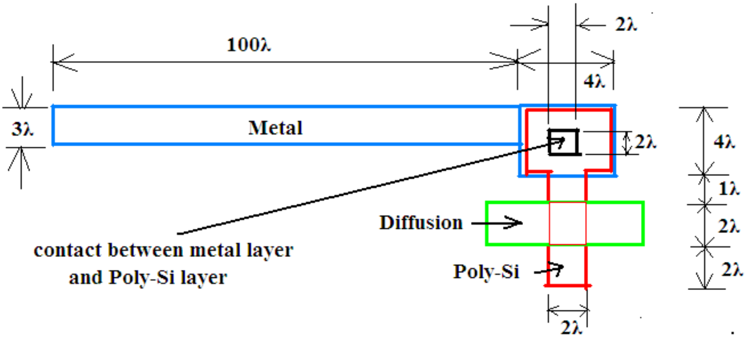


Figure 7.6.2.3. Metal intereconnect to Poly-Si Extension from the Gate of MOS device in 5um Generation Technology.
i. Metal interconnect to substrate capacitance is to be calculated;
ii. Poly-Si to substrate capacitance is to be calculted excluding Gate;
iii. Gate Capacitance is to be calculated;
iv. Total capacitance is to be calculated.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued2

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued2 calculates the charging time constant of the Gate of NMOS through a Pass Transistor.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued2

7.6.3. Charging and discharging delay in CMOS.

Like all the Logic Gates, MOS Logic also experiences propogation delay which results in the degradation of the toggle rate og CMOS gates. In order to minimize this degradation it is important to quantitatively analyze the propogatioon delay.

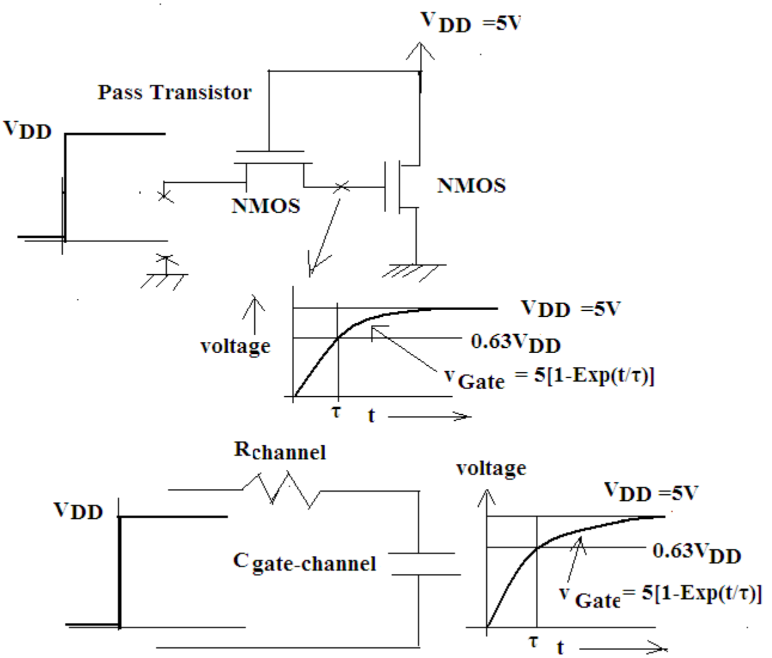


Figure 7.6.3.1. (a)NMOS Gate is being driven through a Pass Transistor.(b)The RC equivalent circuit of the charging path at the Gate of NMOS.

In Figure 7.6.3.1. we have a NMOS being driven through a NMOS Pass Transistor by V_{DD} magnitude Step Voltage where V_{DD} has a magnitude of 5V.

As seen in the Figure 7.6.3.1, a 5V step input results in exponentially growing Gate Voltage. With a time-dealay of $\tau = RC$, Gate Voltage rises to $0.63 \times 5V$. This is evident from the charging equation given in the diagram.

Here $R =$ resistance of the n-channel of pass transistor under ON condition

$=$ Sheet Resistance of the channel $\times (L_{channel}/W_{channel}) = R_{channel}$;

From Table 4.6.1, for 5μm Generation Technology R_{sheet} of n-channel $= 10^4\Omega/\square$ and

$L_{channel} = 2\lambda$ and $W_{channel} = 2\lambda$.

Therefore $R_{chanel} = 10k\Omega \times (2\lambda/2\lambda) = 10k\Omega$;

C = gate-to-channel area capacitance $= C_g =$ area of the gate $\times \square C_{gate\text{-}to\text{-}channel} = 0.01pF$ from the Table 7.6.3 for 5μm Generation.

Hence

For 5μm Generation charging time constant $= R_{chanel} \times C_g = 10k\Omega \times 0.01pF = \tau = 0.1nsec$;

For 2μm Generation charging time constant $= R_{chanel} \times C_g = 20k\Omega \times 0.0032pF$

$= \tau = 0.064nsec$;

For 1.2μm Generation charging time constant $= R_{chanel} \times C_g = 20k\Omega \times 0.0023pF$

$= \tau = 0.046nsec$;

In Table 7.6.3.1 these delays for different generations are tabulated.

Table 7.6.3.1. Delay Time Calculation for 5μm, 2μm and 1.2μm Generations.

Generation	$R_{channel}$	C_{gate}	$\tau_{charging}$	$\tau_{transit}$
5μm	10kΩ	0.01pF	0.1nsec	0.13nsec
2μm	20kΩ	0.0032pF	0.064sec	0.04nsec
1.2μm	20kΩ	0.0023pF	0.046nsec	0.03nsec

In Basic Electrical Properties Chapter 7.3, Equation 7.3.7, the transit time through the channel is given by the equation:

$$\tau_{sd} = \frac{L^2}{\mu_n V_{ds}} \qquad 7.3.7$$

Where L = 5/2/1.2μm, μ_n = 650cm$_2$/(V-sec), V_{ds} = average of V_{DD} = 3V/1.5V/0.75V

By taking the values of the parameters in Equation 7.3.7 we arrive at the transit time tabulated in Table 7.6.3.1.

By inspection of Table 7.6.3.1 we find that with scaling the charging delay decreases and speed performance correspondingly improves.

Since time constant of charging and transit time are nearly equal they are interchangeably used and it is also used as the fundamentl time unit. All timings in a system can be assessed in terms of charging time constant.

To account the variation in the parameters of the devices, we always take the worst case scenario namely 0.3nsec for 5μm Generation, 0.2nsec for 2μm Generation and 0.1nsec for 1.2μm Generation.

We conclude that standard delay unit $= \tau =$ $R\square$(sheet resiatance of the n-channel) $\times \square C_g$;

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 3

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 3 calculates the rise time and the fall timne and proves that for symmetric switching PMOS width must be 2.5 times wider than NMOS width.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 3

7.6.4. Formal Estimation of CMOS Inverter delay.

In Figure 7.6.4.1 we show a CMOS Inverter being driven by square wave and loaded with input capacitance of the subsequent stage. We model the CMOS Inverter during switching transients by constant current source driving the load capacitance.

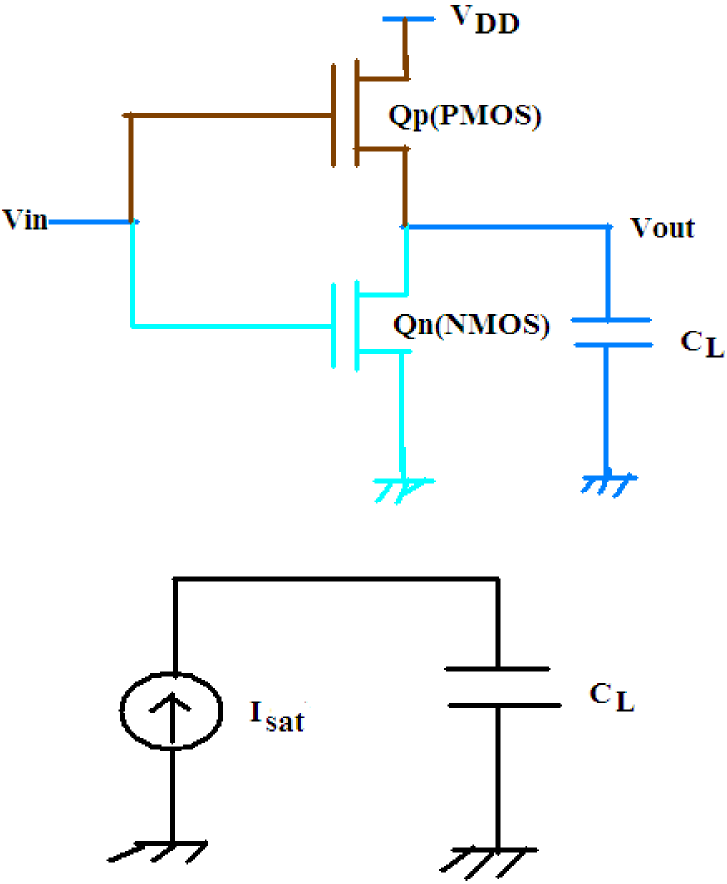


Figure 7.6.4.1. CMOS being driven by square wave and output terminates in a load capacitance.

The switching transient is shown in Figure 7.6.4.2.

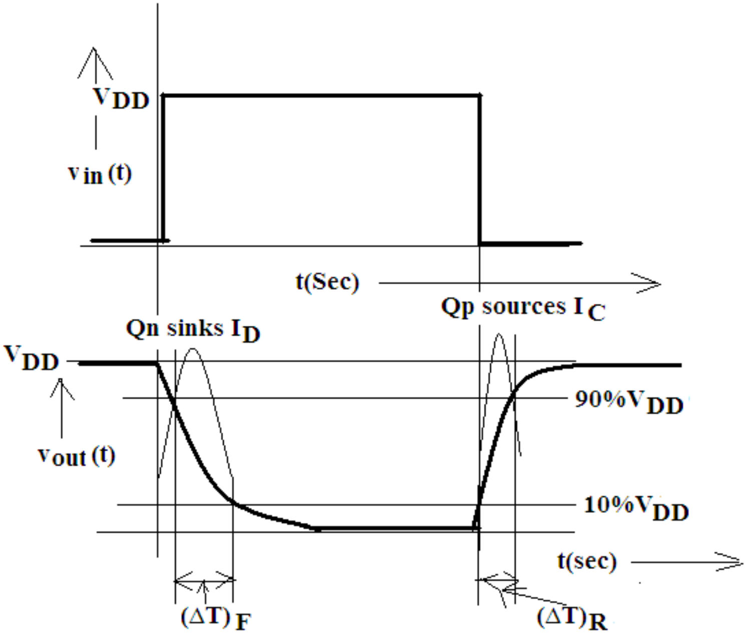


Figure 7.6.4.2. Switching transients of CMOS Logic Gate

As can be seen in Figure 7.6.4.2, during HIGH to LOW switching NMOS (Q_n) provides the low impedance path to ground and hence sinks the discharge current. Hence saturation current of NMOS provides the I_D .

Similarly during LOW to HIGH switching PMOS (Q_P) provides the low impedance path from V_{DD} bus to the load capacitance and hence sources the charging current. Hence saturation current of PMOS provides the I_C .

Referring to the circuit model in Figure 7.6.4.1 we see that:

Rate of change of capacitance voltage = Constant Current/C

Therefore:

$$\frac{dV_C}{dt} = \frac{I_{sat}}{C} \qquad 7.6.4.1$$

Therefore:

$$\Delta V_C = \frac{I_{sat}}{C} \times \Delta T \qquad 7.6.4.2$$

Therefore:

$$\Delta T = \frac{CV_C}{I_{sat}} \qquad 7.6.4.3$$

During output HIGH to LOW transient: NMOS (Q_n) is sinking the current from the load capacitance hence constant current source is a discharge current I_D coming as the saturation current of NMOS where

$$I_D = \frac{\beta_n}{2}(V_{GS} - V_T)^2 \rightarrow fall\ time\ \Delta T_F = \frac{C_L(\Delta V_C)}{\frac{\beta_n}{2}(V_{GS} - V_T)^2} \qquad 7.6.4.4$$

During output LOW to HIGH transient:PMOS(Q_p) is sourcing the current to the load capacitance hence constant current source is a charging current I_C coming as the saturation current of PMOS where

$$I_C = \frac{\beta_p}{2}(V_{GS} - V_T)^2 \rightarrow rise\ time\ \Delta T_R = \frac{C_L(\Delta V_C)}{\frac{\beta_p}{2}(V_{GS} - V_T)^2} \tag{7.6.4.5}$$

But $\Delta V_C = V_{DD}$, $V_{GS} = V_{DD}$ and $V_T = 0.2\ V_{DD}$ therefore

$$\frac{(\Delta V_C)}{(V_{GS} - V_T)^2} = \frac{V_{DD}}{(V_{DD} - 0.2V_{DD})^2} = \frac{1.5625}{V_{DD}} \tag{7.6.4.6}$$

Substituting Eq.7.6.4.6 in Eq 7.6.4.4 and Eq.7.6.4.5 we get:

$$\Delta T_F = \frac{C_L(1.5625)}{\frac{\beta_n}{2} V_{DD}} = \frac{C_L}{V_{DD}} \times \frac{3.125}{\beta_n} \tag{7.6.4.7}$$

$$\Delta T_R = \frac{C_L(1.5625)}{\frac{\beta_p}{2} V_{DD}} = \frac{C_L}{V_{DD}} \times \frac{3.125}{\beta_p} \tag{7.6.4.8}$$

Dividing Eq 7.6.4.8 by Eq.7.6.4.7 we get:

$$\frac{\Delta T_R}{\Delta T_F} = \frac{\beta_n}{\beta_p} \tag{7.6.4.9}$$

From Basic Electrical Properties we know that

$$\frac{\beta_n}{\beta_p} = \frac{\square C_g \mu_n (\frac{W}{L})_n}{\square C_g \mu_p (\frac{W}{L})_p} = [\mu_n (\frac{W}{L})_n] \div [\mu_p (\frac{W}{L})_p] \tag{7.6.4.10}$$

From Device Physics we know that

$$\frac{\mu_n}{\mu_p} = 2.5 \tag{7.6.4.11}$$

We also know from section 7.3.12 titled ‘optimization of geometric parameters to achieve the beswt CMOS performance’ that :

$$[2.5(\frac{W}{L})_n] = \left[\left(\frac{W}{L}\right)_p\right] \tag{7.6.4.12}$$

Substituting the two conditions stated by Eq.7.6.4.8 and Eq.7.6.4.9 in Eq 7.6.4.7 we get

$$\frac{\beta_n}{\beta_p} = 1 \tag{7.6.4.11}$$

Eq 7.6.4.10 implies that we will get symmetric switching transients as well as maximum Noise Margin if satisfy the condition stated in Eq 7.6.4.11 namely

$$W_p = 2.5W_n \tag{7.6.4.12}$$

Hence CMOS layout will be as shown in Figure 7.6.4.3 for optimum performance in all sense of the word i.e in the sense of switching transient and in the sense of noise margin.

Two alternative layouts are shown in Figure 7.6.4.4.

In all the three figures, where metal has to connect to Poly-Si or to n-channel or to p-channel there a contact has to be made symbolized by black cross.

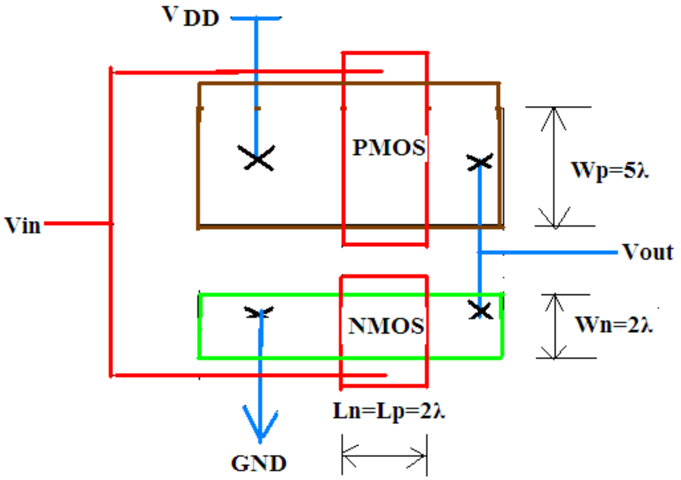


Figure 7.6.4.3. Tha layout of CMOS with Wp=2.5Wn for symmetric switching.
Red is Poly-Si
Brown is PMOS
Green is NMOS
Blue is Metal 1

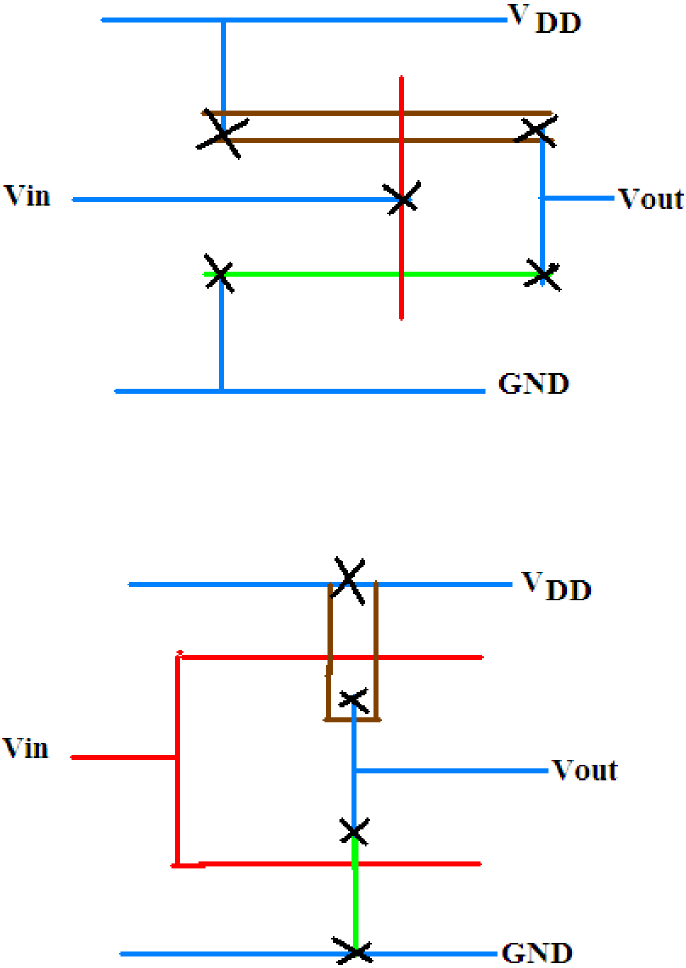


Figure 7.6.4.4. Two alternative stick diagram of CMOS for symmetric switching.
Crosses indicate contact between two different layers.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued4

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued4_module describes the method of driving off-chip circuits with best time response.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 4

7.6.5. Driving off-chip capacitive loads.

We saw in Section 7.6.4 that transient response is:

1. inversely proportional to V_{DD};
2. Directly proportional to load capacitance;
3. Rise-time is 2.5 times Fall-time for equal geometry CMOS .

$$t_f = \frac{C_L \times 3 \times L_n}{V_{DD} C_{ox} \mu_n W_n} \qquad 7.6.5.1$$

$$t_r = \frac{C_L \times 3 \times L_p}{V_{DD} C_{ox} \mu_p W_p} \qquad 7.6.5.2$$

Off-chip capacitances are several orders of magnitude greater than the standard gate-to-channel capacitance □C_g .

Typically off-chip capacitance $C_L \geq$ □C_g . To minimize the charging and discharging delays of the off-chip capacitance we must minimize the pull-up and pull-down channel resistance which can be done by making the channel width W > > the channel L. But this will mean large (W × L) area of the gate of the driver CMOS and this will also cause very large input capacitance of the driver CMOS. The large input capacitance of the driver CMOS will again cause a inordinate delay in the system. So an alternative has to be adopted.

7.6.5.1.Cascaded CMOS Inverters as Drivers.

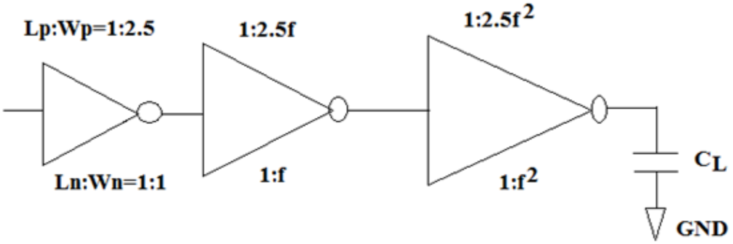


Figure 7.6.5.1.Driving large capacitive loads through a cascade of CMOS inverters

To optimize the drive of off-chip capacitances we must use cascade of CMOS inverters as shown in Figure 7.6.5.1.

We have seen that CMOS inverter driven by another inverter has best performance when Wp =2.5Wn. The proceeding inverters maintain the same pull-up to pull-down resistance ratio but the width is widened by a factor f.

So if first stage has width = Wn = 2λ then the second stage has Wn = 2λ × f and third stage has Wn = 2λ × f × f.

For minimum size lambda based geometries this results in an input capacitance:

$$\begin{aligned} C_{in} &= 1\blacksquare C_g(n-device) + 2.5\blacksquare C_g(p-device) \\ &= 3.5\blacksquare C_g \end{aligned} \tag{7.6.5.3}$$

This optimized CMOS has symmetric switching transients hence :

$$t_r = t_f = 3.5f\tau$$

Therefore total delay per CMOS pair is 7fτ where τ is the standard delay unit as described in Section 7.6.3.

We have to optimize the cascaded configuration so as to minimize the number of stages for a given off-chip capacitance C_L .

Let us define a capacitance ratio y as:

$$y = \frac{C_L}{\blacksquare C_g} = f^N \tag{7.6.5.4}$$

N = number of cascade stages. For a given y we have to arrive at the minimum value of N for a given y.

$$log_{10}y = Nlog_{10}f$$

Therefore

$$N = \frac{log_{10}y}{log_{10}f} \qquad 7.6.5.5$$

The overall total delay in case of CMOS is

$= Nt_r = Nt_f = (N/2)(t_r + t_f) = 3.5Nf\tau.$

By similar consideration total delay in case of NMOS is

$= 2.5Nf\tau.$

So in all cases:

$$Total\ delay \propto Nf\tau = \frac{log_{10}y}{log_{10}f} \times f \times \tau \qquad 7.6.5.6$$

If we tabulate the results of Eq 7.6.5.6 we get Table 7.6.5.1.

Table 7.6.5.1. Tabulation of total delay for various values of f(width multiplication factor) for y = 10 4

$Log_{10}y$	f		Total delay in nsec= K×(logy/logf)×f×τ where τ=0.1nsec
4	1.5		K×0.512 nsec
4	2		K×0.4nsec
4	2.7		K×0.377nsec
4	3		K×0.3784nsec
4	4		K×0.4nsec

So from the Table 7.6.5.1 it is clear that the total delay curve has a minima at f=2.7. Practically we have a soft minima at f=2.7 so we may choose f=3.

In this case from Eq.7.6.5.5, N=logy/logf =8.38 ≈ 8.

So Equation 7.6.5.5 will be used for determining the optimum cascade stage assuming that the width multiplication factor will be 3.

For a positive step input of V DD :

The overall delay for N even: $t_d = 2.5eN\tau$ (NMOS)

$t_d = 3.5eN\tau$ (CMOS)

The overall delay for N odd: $t_d = [2.5(N-1)+1]e\tau$ (NMOS)

$t_d = [3.5(N\text{-}1) + 2]e\tau$ (CMOS)

For a negative step input of V DD :

The overall delay for N even: $t_d = 2.5eN\tau$ (NMOS)

$t_d = 3.5eN\tau$ (CMOS)

The overall delay for N odd: $t_d = [2.5(N\text{-}1) + 4]e\tau$ (NMOS)

$t_d = [3.5(N\text{-}1) + 5]e\tau$ (CMOS)

In the above equations, ‘e’ is the base of the natural log and has a value 2.7

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 5

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 5 deals with the minimization of propogation delay in CMOS logic networks by the use of Super-buffers and BiCMOS drivers.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 5

7.6.6. Super buffers for driving significant capacitive loads.

To drive significant capacitive loads we use super buffers to preserve symmetric switching transients. This considerably minimizes the delay problems. The circuit diagrams of inverting type super buffers is given in Figure 7.6.6.1 and that of non-inverting type in Figure 7.6.6.2.

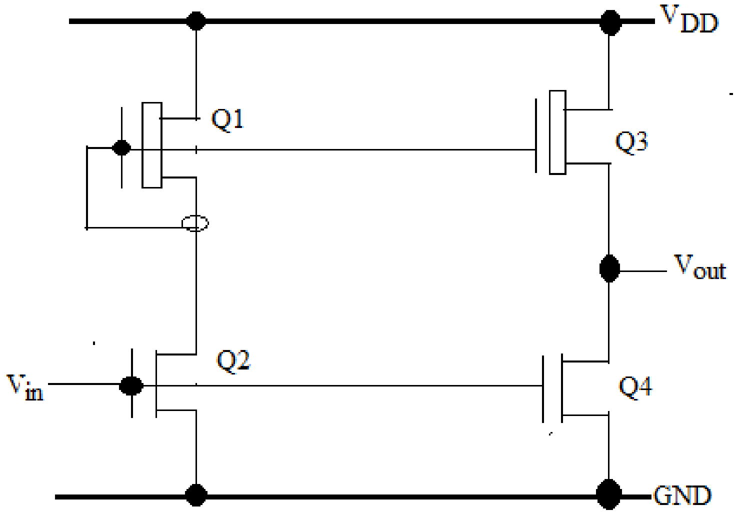


Figure 7.6.6.1. Circuit Diagram for inverting type NMOS super buffers.

In Figure 7.6.6.1. Q1-Q2 comprise (E)NMOS inverter with (D)NMOS acting as pull-up transistor. The output of the inverter drives the gate of Q3 and original input drives the gate of Q4.

When V_{in} = HIGH (V_{DD}), inverter O/P is LOW which pulls down the gate of Q3 to 0V thereby Q3 is turned OFF but Q4 is turned ON. Hence O/P of the super buffer is rapidly pulled down through Q4.

When V_{in} = LOW (0V), inverter O/P is HIGH. Original input turns OFF the pd transistor Q4 but HIGH O/P of the inverter turns ON the pu transistor Q3. Thus O/P of the super buffer is rapidly pulled up through Q3.

Note that in the conventional inverter under pu condition VDD/2 is applied at the gate of pu transistor. In case of super buffer, VDD is being applied to the gate of pu transistor Q3. Hence in the second case pu transistor is being turned ON twice as hard as in the conventional case. Hence pu is twice as rapid. This helps symmetrize the output transients.

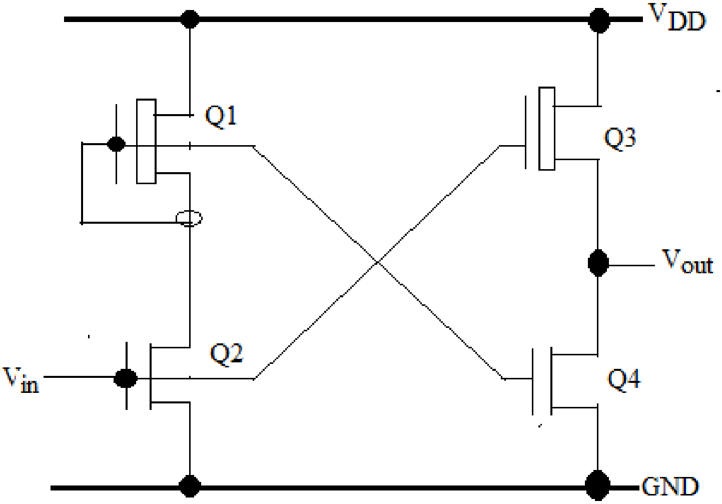


Figure 7.6.6.2. Non-inverting type NMOS Super buffers.

In Figure 7.6.6.2. we have non-inverting type super buffers. It is the same as inverting type except that the inveter output is cross-coupled to Q3-Q4 pair.

Hence when Vin = HIGH, O/P of the buffer is being pulled up through Q3.

When Vin = LOW, O/P is being pulled down through Q4.

Thus Vin = HIGH gives buffer O/P =HIGH and

Vin=LOW gives buffer O/P= LOW.

7.6.7.BICMOS Drivers.

BJT Logic has much better drive capabilities as compared to CMOS Logic.

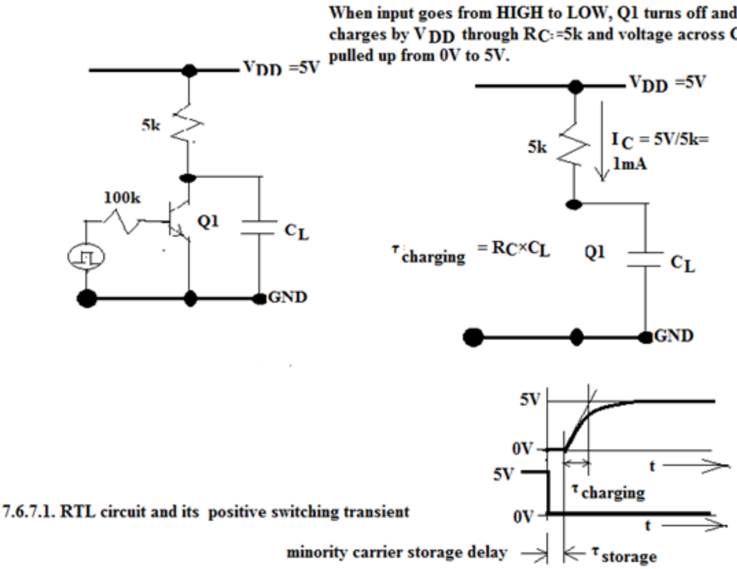
1. BJT has larger transconductance and much larger current per unit area handlng capacity.
2. A much smaller voltage swing is required for switching the BJT transistor hence switching transients are much faster as compared to CMOS circuits.

Because of these reasons BJT drive at the output of CMOS is preferred hence it is called GLUE Logic.

In Figure 7.6.7.1, Resistance-Transistor –Logic (RTL) circuit configuration is shown. Along side the positive switching transient at the output is shown with input experiencing 5V negative step. When I/P goes HIGH to LOW, transistor is first switched OFF. Because the transistor was in saturation therefore it takes time to remove all the stored minority carriers in the base. This causes storage delay ,τ(storage).

Once the transistor is turned OFF, the load capacitance is pulled up through load resistance R_C.

Hence the charging time constant $\tau(\text{charging}) = R_C \times C_L$. The pull-up tramsient with storage delay is shown at the bottom right of the diagram.



When input goes from HIGH to LOW, Q1 turns off and C_L charges by V_{DD} through R_C=5k and voltage across C_L is pulled up from 0V to 5V.

7.6.7.1. RTL circuit and its positive switching transient

In Figure 7.6.7.2 we show the negative transient across the capacitive load at the output of the RTL gate.

The input is HIGH at 5V hence transitor is ON and sink current is caused :

$$Sink\ current = g_m v_\pi$$

Therefore the rate of pull-down across the capacitor is:

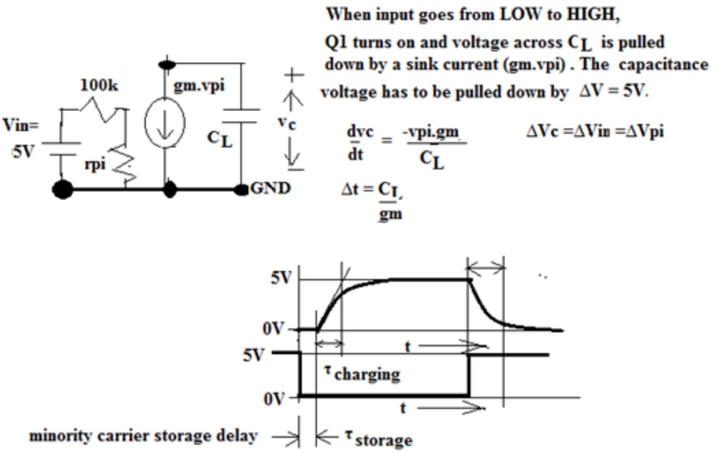
$$\frac{dv_c}{dt} = -\frac{g_m v_\pi}{C_L}$$

This can be rewritten as:

$$\frac{\Delta V_{out}}{\Delta t} = -\frac{g_m}{C_L} \times \Delta V_{in}$$

In Logic Gates :

$$\Delta V_{out} = \Delta V_{in} = 5V$$



When input goes from LOW to HIGH, Q1 turns on and voltage across C_L is pulled down by a sink current (gm.vpi) . The capacitance voltage has to be pulled down by $\Delta V = 5V$.

$$\frac{dvc}{dt} = \frac{-vpi.gm}{C_L} \qquad \Delta Vc = \Delta Vin = \Delta Vpi$$

$$\Delta t = \frac{C_L}{gm}$$

Figure 7.6.7.2. Circuit model of the RTL gate at positive step input. Voltage across the capacitance is pulled down by the sink current (gm.vpi). This give a time dalay of $\Delta t = \frac{C_L}{gm}$

Therefore

$$\Delta t = \frac{C_L}{g_m} = delay\ during\ fall\ time.$$

This clearly shows the delay dependence on transconductance. BJT has a much larger transconductance and therefore delay will be minimized using BICMOS.

So finally the total delay factors in BICMOS are:

Minority carrier storage delay (t1) +

Charging delay or pull up delay +

Discharging delay or pull down delay.

In Figure 7.6.7.3 we make a comparative study of the delays in BiCMOS and CMOS inverters.

Total delay of BiCMOS is given by the following expression:

$$T = T_{in} + \frac{V}{I_d} \times \frac{1}{h_{fe}} \times C_L$$

Where T = total delay; T_{in} = time to charge BE Junction capacitance; hfe = short circuit current gain.

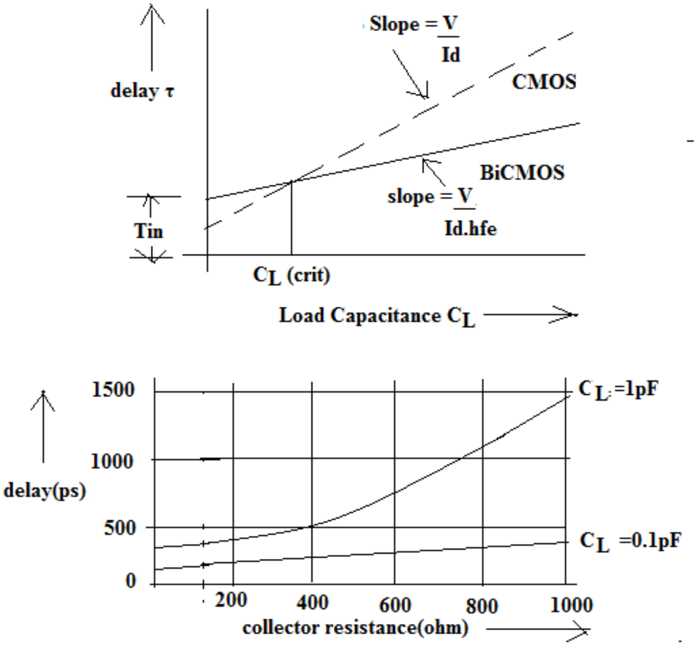


Figure 7.6.7.3. Upper graph is delay estimation in CMOS and BiCMOS logic. Lower graph is gate delay as a function of collector resistance.

In Figure 7.6.7.3 we give two graphs. Upper Graph is the delay performance of BiCMOS and CMOS gates with respect to load capacitance. Lower Graph is the delay performance of BiCMOS with respect to collector resistance Rc for two different values of CL.

From the upper graph it is evident that after a critical value of CL, BiCMOS switching is much faster than CMOS and it gives increasingly better performance as CL increases .

The lower graph shows the necessity of keeping collector resistance low so that the pull-up of capacitive load is rapid. This also is the reason why buried sub-collector layer is included in the BJT structure.

Because BiCMOS can achieve high grade BJT with high gm, high short circuit current gain and low bulk collector resistance hence BJT is included in CMOS fabrication without compromising or overelaborating the basic CMOS process.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 6

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 6 discusses the propogation delay encountered in cascaded pass transistors, along long polysilicon wires and along regular metallic interconnects.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_continued 6.

7.6.8. Propogation delays.

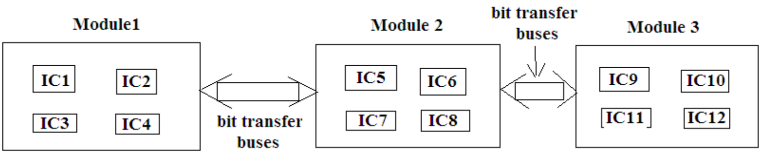


Figure 7.6.8.1. A complete electronic system has several modules.Modules are interconnected through suitable high data transfer buses. Each modules have several IC Chips connected together to fulfil certain system functions. On the IC Chips there are several hundred million components connected through metals or silicides. These interconnects are being stressed to the maximum while transferring data and instructions to and fro.

A typical Electronic System as shown in Figure 7.6.8.1 consists of several modules interconnected through high bit transfer buses. Each module is a PCB which interconnect several ICs through copper tracks etched out on the PCB. Each IC itself has several hundreds of millions of transistors interconnected to achieve a Systen-on-Chip through metal paths or silicide paths.

These metal/silicides interconnects is becoming the biggest bottleneck both cost and performance wise. Scaling of the metal paths pitch in larger and complexICs is leading to long transmission lines with interconnect delay dominating the chip delay as shown in Figure 7.6.8.2.

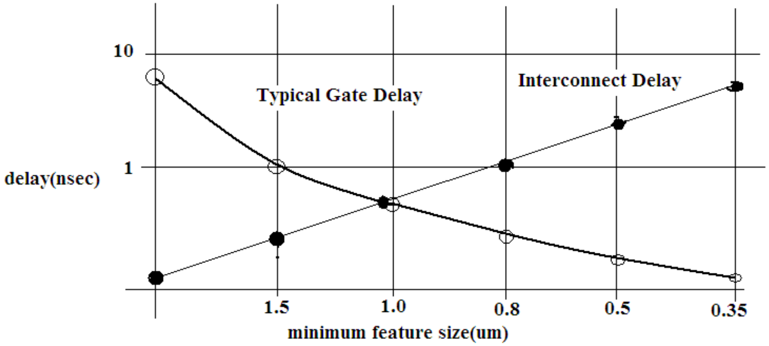


Figure 7.6.8.2. Typical Gate Delay and Interconnect Delay encountered in different generations of technology.

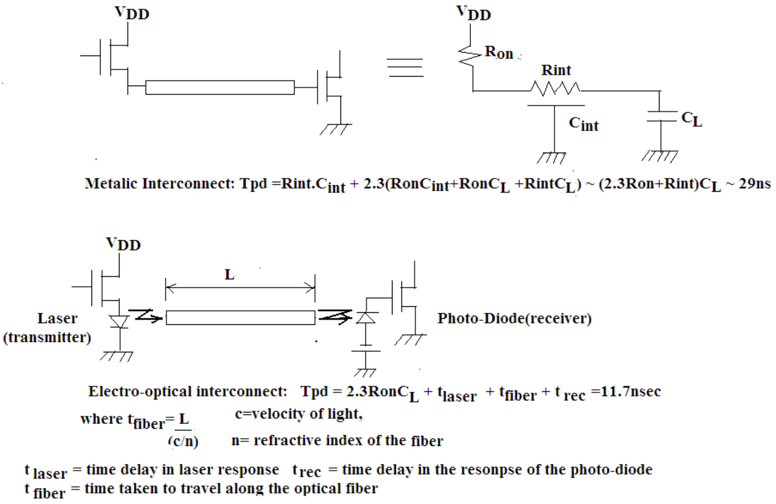
Metallic/Silicides pathways or interconnects on the IC Chip act as the data bus as well as the instruction bus. As the clock speed goes up, these interconnect face skin effect and eddy current losses. Skin effect leads to excessive resistance and heat dissipation. The oscillating signals on the buses induce stray eddy current in the board’s conducting part leading to heavy eddy current losses. With increase in clock rate, attenuation along the copper buses increases exponentially. At 2GHz clock, there is 50% attenuation and at 10GHz, there is 95% attenuation. At few GHz several resonances occur which cause signal to be reflected from the VIAS.

Apart from this, there is severe cross-talk among the metallic interconnects leading to excessive Bit Error Rate (BER). At 10Gb.p.s. rates, cross-talk blurs the signal after 1 meter of ppropogation down the coper bus.

With the increase in the processing speed, the problem of attenuation and cross-talk is becoming more acute.

7.6.8.1.Comparison between metallic and electro-optical interconnects.

In Figure 7.6.8.3 a comparative study is made between metallic interconnect and electro-optical interconnect in terms of propogation delayencountered.



Metalic Interconnect: $T_{pd} = R_{int}.C_{int} + 2.3(R_{on}C_{int}+R_{on}C_L + R_{int}C_L) \sim (2.3R_{on}+R_{int})C_L \sim 29ns$

Electro-optical interconnect: $T_{pd} = 2.3R_{on}C_L + t_{laser} + t_{fiber} + t_{rec} = 11.7nsec$

where $t_{fiber} = \frac{L}{(c/n)}$ c=velocity of light, n= refractive index of the fiber

t_{laser} = time delay in laser response t_{rec} = time delay in the resonpse of the photo-diode

t_{fiber} = time taken to travel along the optical fiber

Figure 7.6.8.2. A comparative study of delays encountered in metallic interconnect and electro -optical interconnect.

In Figure 7.6.8.3 we see that the propogation delay is halfed using electro-optical interconnect as opposed to metallic interconnect.

7.6.8.2. Delay through Cascaded Pass Transistors.

A chain of four cascaded pass transistors and its equivalent distributed circuit model is shown in Figure 7.6.8.3. The gates of the pass transistors are held at VDD so that all the pass transistors are ON.

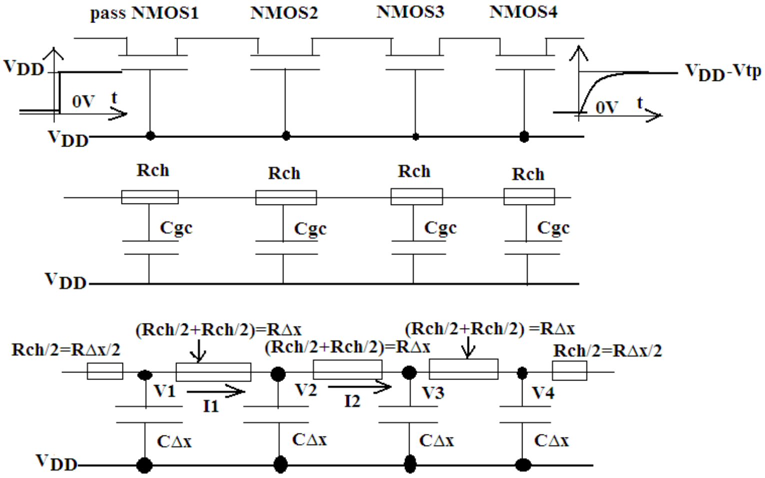


Figure 7.6.8.3. A chain of four cascaded pass transistors with their Gates held at V_{DD} (to allow the passage of signal from the left end of the chain to the right of the chain) and their distributed parameter equivalent circuit.

R=resistance per unit length

C=capacitance per unit length

x= distance along the chain of cascaded pass transistors.

Consider nodes 1,2 and 3.

At node 2 we have:

$$\frac{\Delta I}{C\Delta x} = \frac{dV_2}{dt} \qquad 7.6.8.1$$

But

$$\Delta I = I_1 - I_2 = \frac{V_1 - V_2}{R\Delta x} - \frac{V_2 - V_3}{R\Delta x} = -\frac{\Delta V_1}{R\Delta x} + \frac{\Delta V_2}{R\Delta x} = \frac{1}{R}\left(\frac{\Delta V_2}{\Delta x} - \frac{\Delta V_1}{\Delta x}\right)$$

7.6.8.2

Using Eq 7.6.8.1 and Eq.7.6.8.2

$$\Delta I = C\Delta x \frac{dV_2}{dt} = \frac{1}{R}\left(\frac{\Delta V_2}{\Delta x} - \frac{\Delta V_1}{\Delta x}\right) \qquad 7.6.8.3$$

Rearranging the terms of Eq.7.6.8.3

$$RC\frac{dV_2}{dt} = \frac{\left(\frac{\Delta V_2}{\Delta x} - \frac{\Delta V_1}{\Delta x}\right)}{\Delta x} = \frac{d^2 V}{dx^2}$$

Therefore

$$RC\frac{dV_2}{dt} = \frac{d^2 V}{dx^2} \qquad 7.6.8.4$$

Eq.7.6.8.4 has been derived in Pucknell’s book but this equation is not the standard wave equation. Hence it doesnot have a standard solution and is very difficult to understand as to how Pucknell concludes that the propogation delay = t_{pd} is given as:

$$t_{pd} \propto RC \times N^2 \ (number\ of\ pass\ transistors)$$

In the Book ‘VLSI Design Techniques for Analog & Digital Circuits’ by Randell L.Geiger, Phillip E.Alen and Noel R.Strader, Publisher McGraw Hill Series, 1990, 969 pages,Chapter 7 ‘Basic Digital Building Block’, Section 7.7.1 NMOS Pass Transistors, Page 561, above statement is very clearly inferred by deduction.

The delay for one section is RC.

If second pass transistor is added and only capacitance is considered then delay for two sections will be (R2C);

If second pass transistor is added and second series resistance is considered and capacitance is neglected then delay for two sections will be (2RC);

If while considering two sections 2R and 2C are considered then

Delay is $2R \times 2C = 4RC = 2^2RC$.

If N pass transistors are considered then delay = N^2RC.

Hence in a cascade of pass transistors more than 4 NMOS are never used. If a long chain has to be used then it must be segmented at intervals of 4 NMOS by inverters.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_concluded

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_concluded This module concludes the basic circuit concepts by giving the analysis of the parasitic capacitances associated with fringing-field effect, with interlayer effect,with inter-wire effect,with overlap effect and with peripheral effect.Including these parasitics makes the modelling and simulation more acurate and hence more realistic.

SSPD_Chapter 7_Part 6_Basic Circuit Concepts_concluded

7.6.9. Wiring Capacitances.

We have already discussed area capacitances and its effects on the speed of switching. Here we will discuss the various parasistic capacitances associated with the interconnection wirings. These need to be accounted along with the area capacitances while studying the performance of CMOS circuits. Three main sources of wiring parasitic capacitances are: fringing fields, interlayer capacitances aand peripheral capacitances.

7.6.9.1.Fringing Fields.

In any area parallel-plate capacitance there are transverse field lines as well as fringing field lines which can be significant fraction of the transverse field line as shown in Figure 7.6.9.1. In the standard parallel plate formula (εA/d) which is used for area capacitance calculation we account for transverse field only. To consider the total capacitance we need to calculate the Capacitance due to fringe field.

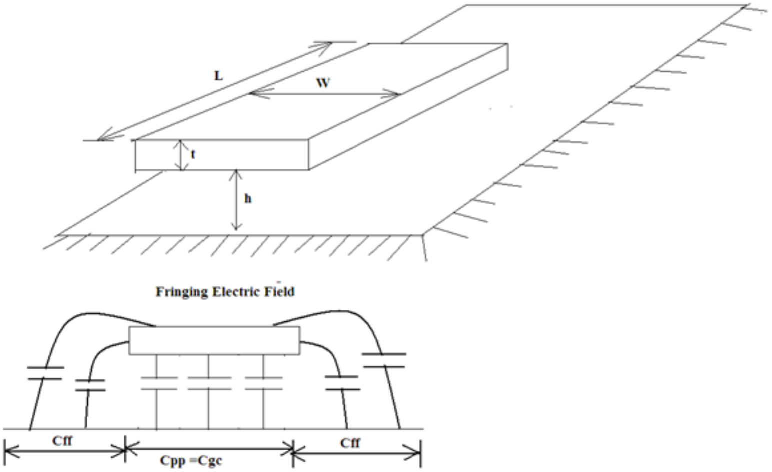


Figure 7.6.9.1. The perspective and the side view of the electric field surrounding Interconnect Segment parallel to the substrate.

$$FF(fringing\ field\ factor) = \frac{C_{Total}}{C_{PP}} \tag{7.6.9.1}$$

Where

$$C_{Total} = C_{pp} + C_{ff}(Capacitance\ due\ to\ fringe\ field\ effect.)$$

It is given:

$$FF = function\left(\frac{t}{h}, \frac{W}{h}, \frac{W}{L}\right)$$

FF is plotted in Figure 7.6.9.2.

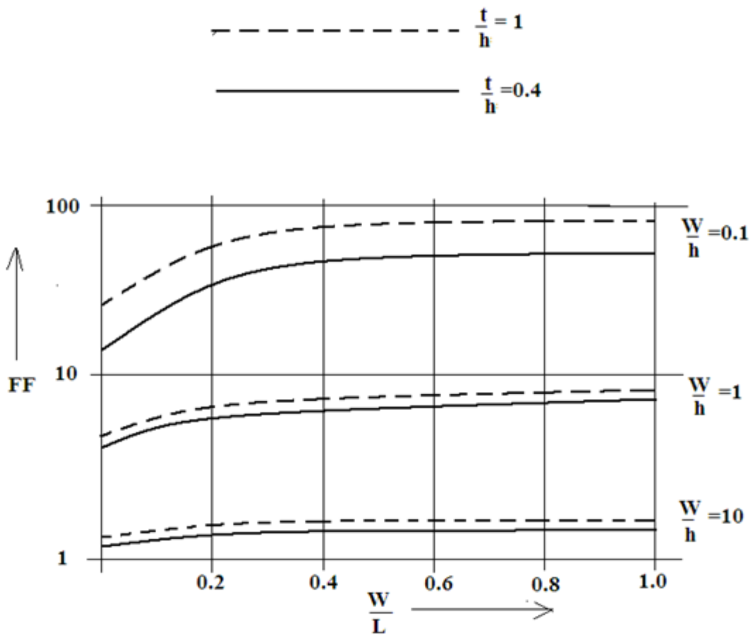


Figure 7.6.9.2. Plot of Fringing_Field_Factor(FF) versus $\frac{W}{L}$ with dependence on $\frac{W}{h}$ and on $\frac{t}{h}$.

From the graph in Figure 7.6.9.2 it is evident that narrower and thicker interconnect segment has a much larger Fringing Field Effect .

Pucknell et.al give the formula for fringing field capacitance for fine line metallization. It is as follows:

$$C_{ff} = \epsilon_{SiO2}\epsilon_0 L\left[\frac{\pi}{\ln\left\{1+\frac{2h}{t}\left(1+\sqrt{\left(1+\frac{t}{h}\right)}\right)\right\}}-\frac{t}{4d}\right] \qquad 7.6.9.2$$

Where L=wire length,

t = thickness of the wire,

h = wire to substrate separation.

7.6.9.2. Interlayer capacitance.

In multi-level metallization IC Chip, interlayer capacitance becomes important therefore in area capacitance chart Table 7.6.2. we have metal 2 to metal 1 and metal 2 to polysilicon. This will occur where two level metal paths overlap or cross each other.

7.6.9.3.Inter-wire Capacitance(due to lateral field).

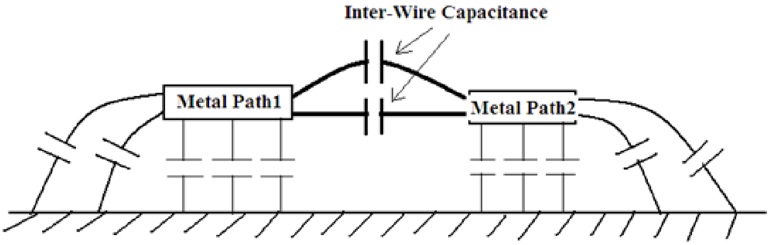


Figure 7.6.9.3. Illustration of inter-wire capacitance due to lateral field.

There is inter-wire capacitance between two metal pathways at the same level due to lateral field as shown in Figure 7.6.9.3. Along side Fringe Field Capacitance and Parallel Plate Capacitance is also shown.

7.6.9.4. Overlap Fringe Field Capacitance.

Two level metal pathways may overlap as shown in Figure 7.6.9.4. In this case Overlap Fringe-Field Capacitance is caused.This has to be ascounted for accurate modeling and simulation.

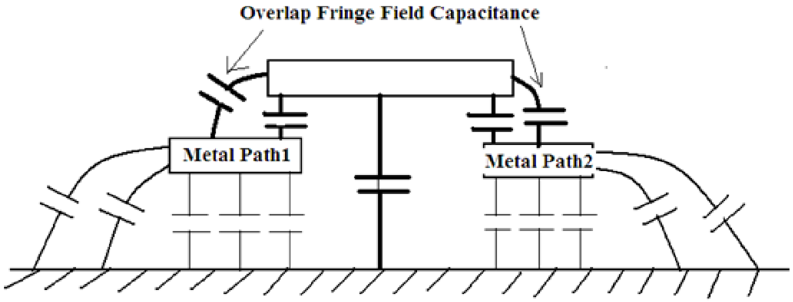


Figure 7.6.9.4. Illustration of Overlap fringe field Capacitance.

7.6.9.5. Peripheral Capacitance.

In Table 7.6.2 we have a row which gives the diffusion(active)region capacitance. These are the junction capacitance of the junction diode made by source and drain diffusion into the Well or into the substrate.

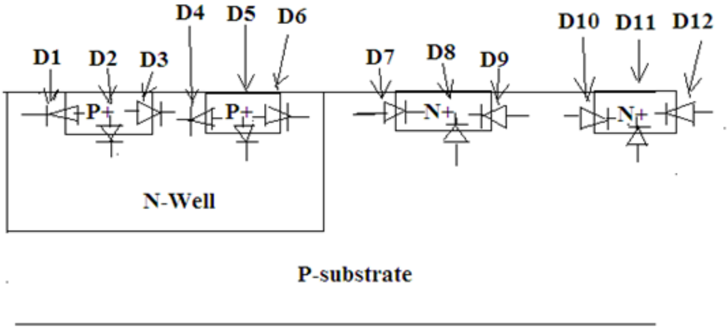


Figure 7.6.9.5. Side-wall(or peripheral) capacitance of the source and drain active region reverse biased diodes with respect to Well or Substrate.

In Figure 7.6.9.5 we show the cross-section of CMOS structure achieved by N-Well Process.

We have P+ Source and P+Drain diffusion into N-Well to realize PMOS and we have N+Source and N +Drain diffusion into P-Substrate to realize NMOS. This created 12 PN Junction Diodes.

D1,D3,D4 and D6 are sidewall diodes and D2 & D5 are the bottom diodes. Table 7.6.2 in the 5th row accounts for the capacitance associated with D2 & D5 only. The side wall diodes junction capacitances are called peripheral capacitances and have not been accounted for.

Similarly N+Source and N+Drain diffusion into P-substrate have created D7,D9,D10 and D12 sidewall diodes and D8 & D11 bottom diodes. Here also 5th Row of Table 7.6.2 acounts for D8 & D11 botom diodes only.Peripheral diodes junction capacitance remain to be accounted for.

As we scale downGeneration after Generation the relative value of peripheral capacitances increases with respect to area capacitances(or bottom capacitances).

If instead of diffusion, the active regions are achieved by ion-implantation then because of the very shallow depths of the active regions the sidewall capacitances are insignificant and hence can be neglected.

Table 7.6.9.1 gives the values of the bottom junction capacitances and sidewall junction capacitances.

Table 7.6.9.1.Typical values for diffusion capacitances.

	Typical values of junction capacitances, bottom and sidewall.		
	5μm	2μm	1.2μm

Area Capacitance(C_a)	1.0×10^{-4}pF/µm^2	1.75×10^{-4}pF/µm^2	3.75×10^{-4}pF/µm^2
Peripheral Capacitance(C_{periph})	8.0×10^{-4}pF/µm^2	Ion-implant	Ion-implant

7.6.9.6. Choice of layers according to the suitability of their functions.

Metal Layers , poly-silicon layers, diffusion layers are chosen according to the functions which they are to perform.

1. V_{DD} and Ground (V_{SS}) should be distributed on metallic paths.
2. Long lengths of poly-silicon layer should not be used because their high sheet resistance. They are suitable only for local interconnects.
3. Capacitive effects are critical in fast signal lines. Fast signal lines should pass over interconnects with low sheet resistance and running over low permittivity dielectric layers because velocity of the signal is:

$$v = \frac{1}{\sqrt{\mu_0 \epsilon_0 \epsilon_r}}$$

Lower is the value of relative permittivity closer is the velocity of propogation to ‘c’(velocity of light in free space)..

Table 7.6.9.2 gives the maximum length of communication lines based on the speed of

switching consideration.

Table 7.6.9.2.Electrical Rules for communication lines.

layer	Maximum legth of communication wire		
	Lambda-based(5μm)	μm-based(2μm)	μm-based(1.2μm)
Metal	Chip wide	Chip wide	Chip wide
Silicide	2000λ	NA	NA
Polysilicon	200λ	400μm	250μm
Diffusion(active)	20λ*	100μm	60μm

*Taking into account of peripheral and area capacitances. NA = not applicable.

Table 7.6.9.3 gives the thickness values of different layers in 0.8μm CMOS process.

Table7.6.9.3.Thickness values of different layers(0.8µm CMOS process)

Layer	Thickness(µm)
Field Oxide	0.52
Gate Oxide	16nm
Poly Si thickness	0.35
Poly-Metal Oxide thickness	0.65
Metal 1 thickness	0.60
Via Oxide thickness	1.00
Metal 2 thickness	1.00
N+ junction depth	0.40
P+ junction depth	0.40
N-Well junction depth	3.5

Different terms used in Table 7.6.9.3 are explained in Figure 7.6.9.6.

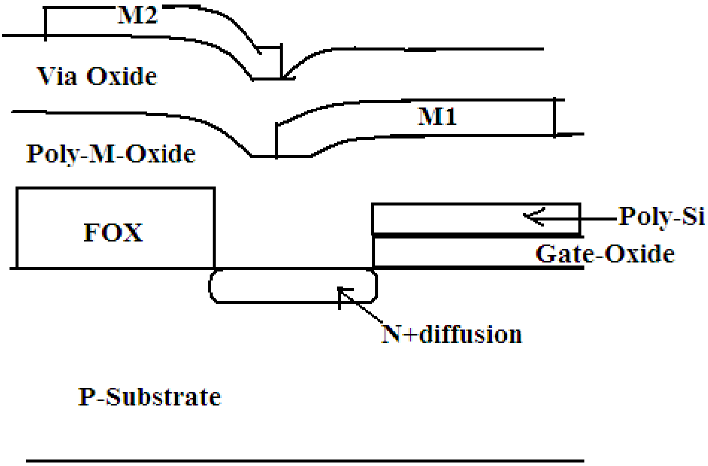


Figure 7.6.9.6. Cross-sectional view of MOS structure with different layers identified.
FOX- Field Oxide

In Table 7.6.9.4. we give the functional suitability of different layers.

Table 7.6.9.4.Functional suitability of different layers.

Layer	R	C	Comments
Metal	low	low	Use for power distribution and global

			signals
Silicide	low	moderate	Modest RC delay. Reasonably long wires are possible. Silicide is used in place of Poly-Si in some NMOS processes
Poly-Si	high	moderate	Moderate RC delay and high IR drop
Diffusion(active)	moderate	high	Moderate IR drop but high C hence hard to drive.

SSPD_Chapter 7_Part 7_Scaling of MOS circuits. SSPD_Chapter 7_Part 7_Scaling of MOS circuits deals with the basic definition of scaling, its effects on the scale of economy in IC production and the various design options it is offering.

SSPD_Chapter 7_Part 7_Scaling of MOS circuits.

7.7.1. What is scaling ?

Scaling is reduction in lateral and vertical dimensions to achieve a higher packing density on Si IC Chip, to achieve faster switching devices, to achieve higher functionality and achieve better scale of economy.

In 1966 emulsion mask patterns were contact printed using UV light on 200mm wafer with the smallest feature size of 25μm.

In 2007 using Deep Ultra Violet(DUV)193nm immersion Lithography, 30nm ± 6 nm feature size are imprinted on 300mm wafer.

In 2012 using Extreme Ultra Violet Light(EUV)100nm reflective mirror exposure system (Next Generation Lithography-NGL), less than 30nm feature size is expected to be imprinted on 450mm wafer.

In 70s growth of Personal; Computer fueled the

CMOS scaling.

Today in 21^{st} century, Internet, high speed communication that converges data, video and audio and mobile communication is fueling CMOS scaling consistent with Moore’s Law and has sharply brought in focus the need for cheap, high speed and low power devices for which Si:Ge is emerging as the the most promising candidate.

DSP and ‘DSP & Real World Interface’ are at the heart of Internet, so developing System-on-Chip(SOC) and integration of digital and Analog/RF function became high priority area of research. Advanced SiGe BJT Process also shrinks the lateral and vertical dimensions and boosts the performance hence it has become the technology of choice for 24GHz radar for blindside detection, for 77GHz radar system automobile collision warning or advance cruise control, for 60GHz Wi-Fi chips for next generation wireless LAN and backbone networks, for software defined radios, for cellular handset and for high frequency automatic test equipments.

7.7.2.How does scaling help achieve greater functionality ?

As we downscale the lateral and vertical dimensions we simultaneously add more functions to the chips.

In 1971 μP4004 was a 16-pin DIP packaged chip which had the CPU built in the die. It was a LSI chip which along with three other LSI chips could be used to build a complete functional computer. These LSI chips were : 4001(ROM-4-bit output), 4002(RAM-4-bit input/output) and 4003 static shft register for expanding input/output lines. The photograph of the exposed 16-pin μP4004 chip and its hermetically shielded picture are shown in Figure 7.7.1.

4004 LSI chip used 10μm PMOS(E) technology. It had 2300 transistors and it could carry out 92,000 instructions per second that is 10.8μsecond per instruction.

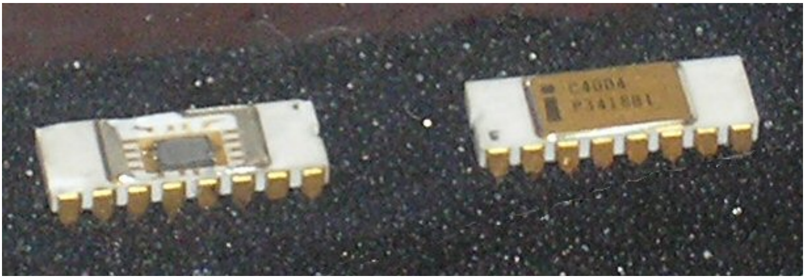


Figure 7.7.1. Photograph of μP4004,16-Pin DIP packaged chip. Left is the exposed chip. Right is the hermetically sealed chip.

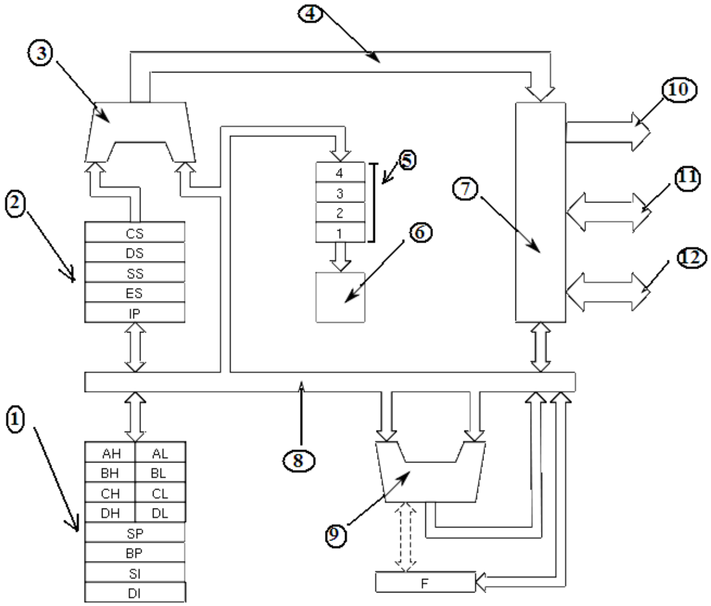


Figure 7.7.2. Block diagram of the Intel 8086 microprocessor

[1. Block of general purpose registers, 2 Block segment registers, 3 20 BIT combiner, 4 Internal bus C, 5 Queue commands, 6 The control system, 7 The control system bus, 8 Internal Bus A, 9 Arithmetic logic unit (ALU), 10 Address bus, 11 Data bus, 12 Rail Control F. Registry tags, AX -accumulator , BX - register base CX - counting register, DX - data register, SP - stack pointer, BP - base pointer, SI - source index, DI - Destination Index ,CS – code. Segment DS - data segment, SS - stack segment, ES - extra segment, IP - Instruction Pointer.]

In mid-70s, all the parts of CPU were put on one die for µP 8086 (16-bit processor, 40 pin DIP package) as shown in Figure 7.7.2.

In 80s µP80286 and µP80386 were introduced which had ×87 Floating-point hardware on a separate chip on the same mother board.

In 1989 , tightly pipe-lined 8086 architecture was introduced as µP80486. This included ×87 Floating-point hardware on-die.

In 1990s, Single Instruction Multiple Data (SIMD) started out as a separate chip. SUN Microsystyem moved it on-chip when VIS was introduced to UltraSPARC.

In 1992-93 Pentium was introduced which had CPU + Memory + Input/Output port built on the same chip plus it had a second ALU. It was a Intel’s first superscalar processor.

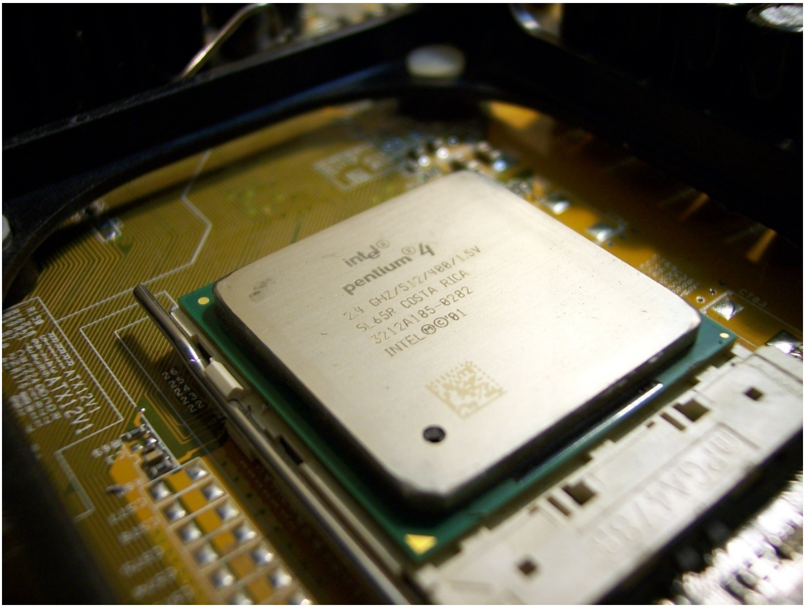


Figure 7.7.3. Photograph of hermetically sealed Pentium 4 chip with 423 and 478-pin PGA packages.

L1 cache was moved on-die in Pentium 2 Architecture.

Latest Pentium architecture started out with a 400 MHz system bus and 256KB L2 cache (later increased to 800 MHz and 2MB). The first models contained 42 million transistors, used the 0.18 micron process and came in 423-pin and 478-pin PGA packages. Intel's first Pentium 4 chipset was the 850 and supported only Rambus memory (RDRAM), but subsequent chipsets switched to DDR SDRAM.

Subsequently Floating-point hardware moved on-die.

Thus new features were added in scaled down chip to add performance-enhancing functionality to a processor die.

7.7.3. Cost Metrics of IC Chip.

$$\mathit{Cost\ of\ a\ processor\ (IC\ chip)} = \frac{\mathit{Die\ Cost} + \mathit{Cost\ of\ Testing\ Die} + \mathit{Cost\ of\ packaging\ and\ final\ test}}{\mathit{Final\ Test\ Yield}} \qquad 7.7.1$$

Final Test yield is the % of successful dies produced hence it is unity or less.

The die cost is calculated by the formula:

$$\mathit{Die\ cost} = \frac{\mathit{Wafer\ cost}}{\frac{\mathit{dies}}{\mathit{wafer}} \times \mathit{yield}} \qquad 7.7.2$$

The denominator of Eq.7.7.2 gives the the total number of good dies or expected number of working dies achieved on a given wafer. It is the product of total number of dies achieved on a wafer and the yield of the given wafer.

Dies per wafer is calculated by the following formula:

$$\frac{Dies}{Wafer} = \pi D\left[\frac{D}{4S} - \frac{1}{\sqrt{2S}}\right] \qquad 7.7.3$$

Where D = wafer diameter in cm and S = die area in cm2.

Rewriting Eq.7.7.3 we get:

$$\frac{Dies}{Wafer} = \left[\frac{\pi\left(\frac{D}{2}\right)^2}{S} - \frac{\pi D}{\sqrt{2S}}\right] \qquad 7.7.3a$$

Eq.7.7.3a can be interpreted as:

$$\frac{Dies}{Wafer} = \#of\ gross\ dies - \#of\ imperfect\ dies. \qquad 7.7.3.b$$

Yield of the dies from a wafer is calculated by the following formula:

$$Die\ yield = Wafer\ yield \times [1 + \frac{\frac{defect}{area} \times die\ area}{\alpha}]^{-\alpha} \qquad 7.7.4$$

Where wafer yield is generally 100% meaning wafer is perfect,

defect/area is given as defects per cm2 and die area in cm2 is to be used.

There are 3 kinds of defects: random, systematic and parametric. Here we are concerned with random defects which is a measure of random manufacturing defects on the wafer.

α = empirical parameter that corresponds to the critical masking levels, a measure of manufacturing complexity of the circuit.

Using Eq 7.7.4 and Eq.7.7.3 in conjunction with Eq.7.7.2, the die cost is calculated for seven processor chips as given in Table 7.7.1.

Table 7.7.1.Die Cost of various processors.

Chip	Wafer dia. (mm)	WaferCost($)	Defects per cm2	Die Area(mm2)	Diesperwafer	Yield(%)	DieCost($)
386DX	150	900	1	43	360	71	4
486DX2	150	1200	1	81	181	54	12
PowerPC601	150	1700	1.3	121	115	28	53
HPPA7100	150	1300	1	196	66	27	73
DECα	150	1500	1.2	234	53	19	149
SuperSPARC	150	1700	1.6	256	48	13	272
Pentium	150	1500	1.5	296	40	9	417

Using Eq.7.7.1 we calculate the cost of the seven processors as given in Table 7.7.2.

Table 7.7.2. IC Chip variable Costs.

Chip	Die cost($)	Package			Test & Assembly Cost($)	Total cost($)
		pins	type	Cost($)		
386DX	4	132	QFP	1	4	9
486DX2	12	168	PGA	11	12	35
PowerPC601	53	304	QFP	3	21	77
HPPt7100	73	504	PGA	35	16	124
DECα	149	431	PGA	30	23	202
SuperSPARC	272	293	PGA	20	34	326
Pentium	417	273	PGA	19	37	473

As we see that with increasing complexity and increasing number of functions per chip, die area is bound to increase as it has in 5th column of Table 7.7.1. Increased die area means reduced die yield which will mean increased die cost hence chip cost.

The manufacturing process dictates the wafer cost, wafer yield, and defect density So, the sole control of the designer is the die area – how much functionality should be packed into a chip for it to be “the most cost effective”?

Empirically cost per die grows roughly as the square of the die area. Infact it is a non-linear function of die area.

7.7.4. Cost metrics dependence on Die Area, on Defect Density and on Alpha(circuit manufacturing complexity parameter)

In 2006, for 90nm Generation Devices ‘Alpha 21264C’ fabricated on 300mm Wafer,

the density of defects = 0.5per cm^2 and alpha, α, = 4.

The cost of 300mm Wafer was $4700 and wafer yield was 95%.

Using Eq.7.7.3, dies/wafer = 231.

Using Eq.7.7.4, die yield = 0.555

Therefore Good dies/wafer = 128.

$$Cost\ per\ good\ die = \frac{\$4700}{128} = \$36.72$$

‘Alpha 21264C’ chip has 524-pin CLGA package.

$$Package\ cost = \$25$$

$$\text{Cost of testing the die} = \frac{\text{Cost of testing hour} \times \text{Average die test tiime}}{\text{Die yield}}$$

$$= \frac{\$400 \times \frac{25\ seconds}{3600sec/hr}}{0.555} = \$5.5$$

Using 7.7.1

$$\text{Cost of Alpha } 21264C = \frac{\$36.72 + \$5.5 + \$25}{100\%} = \$67.22$$

By a similar algorithm, the total variable costs of 5 Processors are calculated and tabulated in Table.7.7.3.

Table 7.7.3. Variable costs of 5 Processors using 90nmGeneration Technology and 300mm Wafer. (Part I)

μP	Die area(mm2)	Pin	Wafer cost($)	Package	Dies/ Wafer	Die yield
Alpha21264C	115	524	4700	CLGA	231	0.555
Power 3 II	163	1088	4000	SLC	157	0.452
Itanium	300	418	4910	PLC	79	0.266
MIPS	204	527	3700	CPGA	122	0.383

R14000						
Ultra SPARCIII	210	1368	5200	FC-LGA	118	0.374

Table 7.7.3. Variable costs of 5 Processors using 90nmGeneration Technology and 300mm Wafer. (Part II)

Functional dies/ wafer	Die cost($)	Test cost($)	Package Cost($)	Total cost($)
128	36.72	5.5	25	67.22
71	56.34	5.16	20	81.50
20	245	12.54	20	277.54
46	80.43	7.98	25	113.41
44	118.18	10.7	30	158.88

As we have already noted and it is further verified from Table 7.7.3 that total cost is a nonlinear function of the die area. Generally it is square of the die area.

7.7.4.1. Effect of the Defect density on the cost metrics.

In Table 7.7.4, the effect of defect density on the cost metrics is evaluated.

Table 7.7.4.Total cost of a processor ITANIUM (die area = 3cm 2) versus defect density with Wafer diameter of 300mm.

Defect density(d/cm2)	D/W	Y	FD/W	DC($)	TC($)	Package Cost($)	Total cost($)
0.3	79	0.422	33	148.8	7.9	20	176.39
1	79	0.101	8	612.5	32.91	20	665.41

D/W – dies per wafer, Y- die yield, FD/W – functional dies per wafer, DC- die cost, TC- test cost.

As is evident from Table 7.7.4, increase in defect density has drastic reduction effect on the yield and hence it has drastic enhancement effect on the die cost. So defect density has to be kept under tight control and it needs to be reduced as die area is increased if the die yield has to be maintained at a constant level.

7.7.4.2. Effect of Alpha (manfactruring complexity parameter) on Cost Metrics.

In Table 7.7.5 the effect of Alpha parameter on the cost metrics of Alpha21264C (die

area = 1.15cm^2) processor is studied.

Table 7.7.5. Total cost of a processor Alpha21264C(die area = 1.15cm^2) versus alpha parameter.

Alpha parameter	D/W	Y	FD/W	DC($)	TC(S)	Package Cost($)	Total cost($)
α = 4	231	0.415	95	49.47	7.36	25	81.44
α = 6	231	0.404	93	50.54	7.57	25	83.11

Increase in levels of interconnection causes only a slight reduction in yield and hence a slight increase in total cost. This means that improved routability of the signal in the IC Chip is inexpensive.

7.7.5. Wafer size transition from 200mm to 300mm and now to 450mm.

IC Industry's ability to increased productivity by 25 to 30% per year is the combined result of wafer size transitions, shrinking device geometries, equipment productivity improvement and incremental yield improvements. Wafer size transitions historically account for 4% out of the stated 25-30% productivity increase. This is precisely why wafer size has been increasing periodically as shown in Table 7.7.6.

Table 7.7.6. Wafer size transition leading to increased throughput and increased productivity and decreased unit cost.

Year	WD(mm)	t(µm)	WC($)	D/cm2	DA(mm2)	D/W	Yield	FD per wafer	Die cost($)
59-61	25	200	?	1	81	6	48%	3	
61-63	51	275	?	1	121	7	36%	3	
63-66	76	375	?	1	196	11	22%	3	
66-68	100	525	?	1	234	19	17%	3	
68-70	130	625	?	1	256	34	16%	5	
70-76	150	675	1491	1	296	40	13%	5	298
76-00	200	725	1050	1	296	80	13%	10	105
00-12	300	775	735	1	296	200	13%	26	28
12-?	450	925	515	1	296	479	13%	62	8

WD-Wafer diameter, WC-Wafer cost, D-defect,DA-die area , D/W- gross dies per wafer,FD- functional die.

Starting wafer cost per unit would be reduced by

30% under wafer size increase according to VLSI Research from 1970 onward.

Table 7.7.6 clearly indicates that with increase in Wafer size the scale of economy of IC production improves. With constant die area, yield remains constant but if functionality increase leads to increased die area then defect density will have to be brought down concomitantly so that the product of defect density and die area is constant to ensure constant yield . The constantancy in the die yield only will ensure improved scale of economy.

Driving force for all wafer size transitions include factors of ever increasing die size and increasing number of integrated functions per chip.

Trends indicate that wafer size transitions industry-wide have typically enabled 4% per year productivity improvement and transition to 300mm wafer size should provide 2 to 4% per year lower IC cost per cm^2 .

7.7.5.1. Hydrogen injection process to keep the defect density at the lowest with increase in wafer size.

Hydrogen injection during wafer manufacturing has helped in the following manners:

1. Improved uniformity across the wafer, across the run and from run to run;

2. Control of defects such as oxidation induced stacking faults and control of point defects simultaneous process.

Hydrogen injection produces higher yields helping to improve cost effectiveness.

7.7.6. Design options offered by Scaling and Wafer size transition..

Scaling gives us a higher transistor budget meaning higher transistor density. So for same functionality die area can be decreased and for same die area functionality can be increased.

For same die area, increased functionality means more transistors and hence more features. This will pay off in terms of increased CPU performance. This is called addition of performance-enhancing functionality to a processor die.

In the second option where we go for the same functionality and reduced die area we have four dividends:

1. we get a higher yield and hence decreased die cost;
2. reduced power dissipation because of reduced silicon area;.
3. higher clock rate if power dissipation is maintained at original level;
4. Or a combination of higher clock rate and

reduced power dissipation.

Wafer size increases and concomitant defect density reduction allows the yield to be maintained constant even with increased die area. Thus number of functional chips per wafer are kept constant though much larger in size and much larger in performance enhancing functionality.

Processor's power consumption and dissipation has been on the increase along with die size. This is creating the limitation of Fire Wall (P_0 Watts = power dissipation density × die area) , the maximum limit of dissipation which can be tolerated by a silicon chip.

Here again we have two options:

1. Multi-core procesors which is a combination of smaller , cheaper and less power hungry processor working in parallel to do the same job or
2. Single core processor which is larger, more expemsive and more power hungry.

Manufacturer's are opting for multi-core processors.

TodayMobile Computing demands that more functionality be added to a single die but not for the purpose of increased performance buit for an ideal mobile computing environment. This will mean combining non-volatile memory, volatile memory, CPU, multiple types of wireless capabilities (blue tooth, 802.11b, 802.11g, GSM e.t.c) on a single die. Here again we can integrate them on ta single die or combine multiple chips with multiple functions into a single module.

SSPD_Chapter 7_Part 7_Scaling of MOS circuits_continued2

SSPD_Chapter 7_Part 7_Scaling of MOS circuits_continued2 deals with the scaling models with constant field, constant voltage and dimensions and voltage being scaled by different scaling factors.

SSPD_Chapter 7_Part 7_Scaling of MOS circuits_continued2

7.7.7. Scaling models and scaling factors.

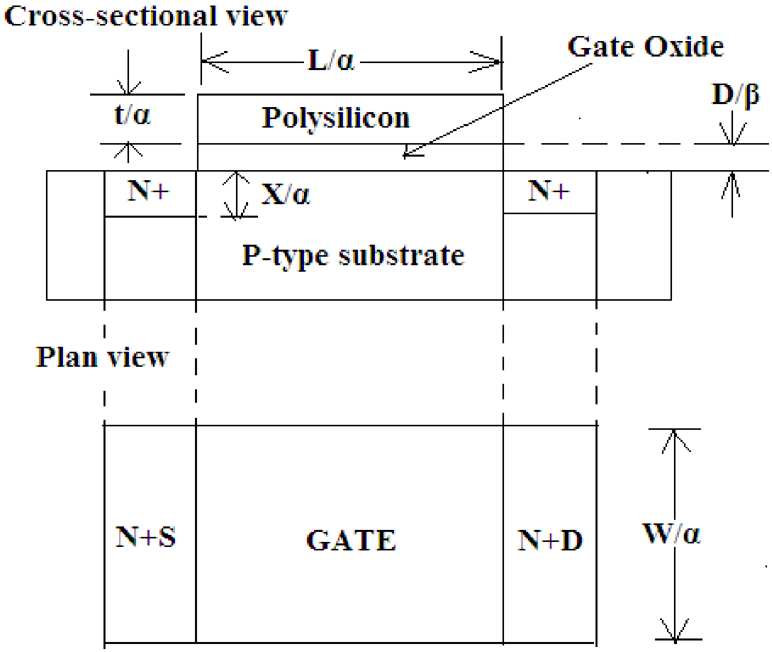


Figure 7.7.7.1 Cross-sectional view and the top view a NMOS Transistor.

In Figure 7.7.7.1. NMOS cross-secctional view and top view are shown.

Linear dimensions both vertical and lateral are being scaled by 1/α..

DC Biasing voltage ,V_{DD} , and Oxide thickness,D, are being scaled by 1/β.

We have three scaling models:

1. Constant Field Scaling Model- here physical dimensions and bias voltage have the same scaling factor therefore $\alpha = \beta$;
2. Constant Voltage Scaling Model- here physical dimensions are scaled but bias voltage is kept constant therefore $\beta = 1$;
3. General Scaling Model- here the two scaling factors are kept different.This is more realistic and presently widely practiced in the Industries.

7.7.7.1.Scaling Factors for Device Geometrical parameters

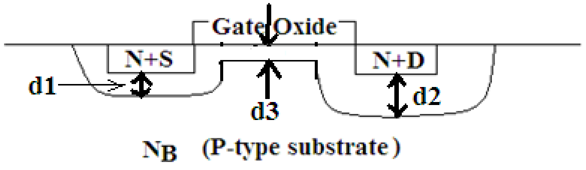
1. Gate Area = $A_g = L \times W$. Each dimension is scaled by 1/α hence A_g is scaled by $1/\alpha^2$.

2.Gate Capacitance per unit area = $C_{OX} = \varepsilon_0\varepsilon_{OX}/D$ where D is scaled by 1/β therefore C_{OX} is scaled by β.

3. Gate Capacitance , C_g , $= C_{OX} \times LW$ has a scaling factor of β/α_2 .

4. Parasitic Capacitance, $C_X = \varepsilon_0 \varepsilon_{Si} A_X/d =$ junction capacitance of the diode at Source-Substrate Junction and at Drain-Substrate Junction. ‘d’ is the depletion width of the diode. Since Source and Drain are at 0V and 5V respectively and substrate is also at 0V therefore depletion region around Source and Drain are d1 and d2 respectively extending primarily into P-Substrate because the two diodes are one-sided step junctions There is a depletion region d3 under the gate in the substrate. This is due to the trapped charges in the gate-oxide and due to the positive voltage a the gate electrode.

i. illustration of depletion width



At Source-Substrate and Drain-Substrate there are N+P junction diodes with corresponding depletion regions d1 and2 respectively. Similarly there is a depletion region under the gate due to trapped charges in the oxide layer and due to positive volatge at the gate.

ii. Graphical plot of 'd'(um) and built-in Electric Field E(V/cm) versus Background doping of the substrate

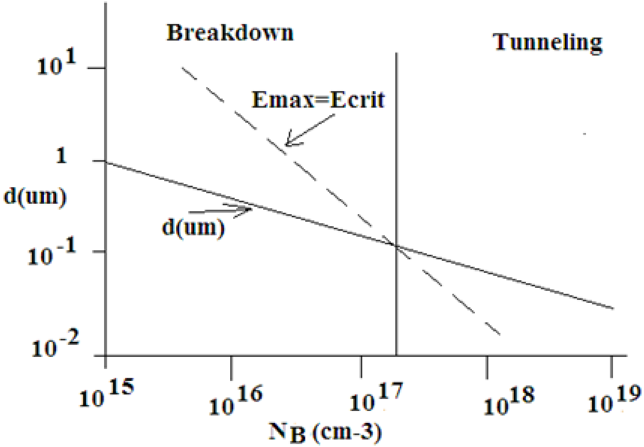


Figure 7.7.7.2. i.Depletion width d1,d2 & d3 around the source, drain and gate.
ii. Plot of 'd' with respect to background doping N_B .

In Figure 7.7.7.1 cross-sectional view of NMOS structure with depletion region of ‘d’µm width is illustrated in the upper figure and in the figure below ‘d’µm’ plot with respect to Background doping of P-substrate is shown.

From Device Physics we know that for one-sided step junction:

$$d = \sqrt{\frac{2\epsilon_0\epsilon_{Si}}{qN_B}}\sqrt{\varphi_{B0} + V_{DD}} \qquad 7.7.7.1.1$$

As the channel length is scaled down, depletion region width ‘d’ must also be scaled down to prevent the source and drain depletion regions from meeting. This is probably to prevent “Punch Through” - a phenomena which occurs in BJT also.

During scaling, cross-sectuional area A_X is scaled by $1/\alpha^2$ and Backgound Doping is increased so that d is scaled by $1/\alpha$.

So the scaling factor of C_X is :

$$\frac{1}{\alpha^2} \times \frac{1}{\frac{1}{\alpha}} = \frac{1}{\alpha} \qquad 7.7.7.2$$

Increase in substrate doping N_B decreases ‘d’ but increases threshold voltage V_{Th} which is against the required trend of scaling and hence cannot be permitted therefore substrate doping is selectively increased around Source and Drain but not under the Gate by the application of deep channel implantation.

7.7.7.2.Scaling factors for different performance parameters for General Scaling Model.

i.Carrier density in the channel = Q_{ON} = $C_{OX} \times V_{GS}$. (since V_{GS} = V_{DD} hence scaling factor for V_{GS} is the same as for V_{DD}). Hence scaling factor for Q_{ON} = $\beta \times (1/\beta)$ = 1

ii. Channel resistance R_{ON} is given as follows:

$$R_{ON} = \rho \times \frac{L}{W \times \Delta} = \frac{1}{\sigma} \times \frac{L}{W \times \Delta} \qquad 7.7.2.1$$

Where ρ is resistivity of the ON channel and σ is the conductivity of the ON channel and

$$\sigma = \mu_n q n_n \qquad 7.7.2.2$$

Where μ_n = drift mobility of electrons since we are dealing with NMOS,

q= charge of an electron,

n_n = number density or electron number per unit volume of the channel therefore

$$q n_n = \frac{Q_n}{\blacksquare \times \Delta} = \frac{surface\ charge\ density}{\Delta} = \frac{Q_{ON}}{\Delta} \qquad 7.7.2.3$$

Substituting Eq.7.7.2.2 and Eq 7.7.2.3 in Eq.7.7.2.1 we get:

$$R_{ON} = \frac{L}{W \times \Delta} \times \frac{1}{\mu_n \times \frac{Q_{ON}}{\Delta}} = \frac{1}{\mu_n Q_{ON}} \times \frac{L}{W} \qquad 7.7.2.4$$

Therefore From Eq.7.7.2.4: $R_{ON} = (L/W) \times [1/(\mu_n Q_{ON})]$

Mobility is unaffected by scaling. Hence scaling factor is $1 \times 1 = 1$

iii Gate Delay $\tau_d = C_g R_{ON} = (L \times W) C_{OX} \times (L/W) \times [1/(\mu_n Q_{ON})] = L^2 C_{OX} [1/(\mu_n Q_{ON})]$

Therefore scaling factor of Gate Delay is $(1/\alpha^2) \times \beta \times 1 = \beta/\alpha^2$;

iv Maximum toggle rate $= f_{max} = 1/\tau_d$. Therefore the scaling factor is the reciprocal of that of Gate Delay i.e. (α^2/β) .

v Saturation Current $I_{DSS} = (\mu_n C_{OX}/2)(W/L)(V_{GS}-V_{th})^2$ Here both V_{GS} and V_{th} are being scaled by $(1/\beta)$. Hence the scaling factor is $\beta \times 1 \times (1/\beta)^2 = 1/\beta$.

vi. Switching energy per gate $E_{gate} = (1/2) C_g (V_{DD})^2$. The scaling factor is $(\beta/\alpha^2)(1/\beta)^2 =$

$1/(\alpha^2 \beta)$.

vii. Power dissipation per gate P_{gate} = Quiescent Power + Dynamic Power

$= (V_{DD})^2/R_{ON} + f_{max} \times E_{gate}$. Hence the scaling factor is $(1/\beta)^2 \times 1 + (\alpha^2/\beta) \times (1/(\alpha^2 \beta)$

$= 1/\beta^2$.

viii. Power dissipation per unit area = $P_{\square} = P_{gate}/A_{gate}$. Therefore the scaling factor = $(1/\beta^2)/(1/\alpha^2) = (\alpha/\beta)^2$.

ix. Power Speed Product is the Figure of Merit of the Gate = $P_{gate} \times \tau_d$. This has a scaling factor = $1/\beta^2 \times \beta/\alpha^2 = 1/(\beta\alpha^2)$.

Table 7.7.7.2.1. Scaling factors for different performance parameters for General Scaling Model.

Performance parameter	Formula	Scaling Factor
Q_{ON}	$C_{OX} \times V_{GS}$	1
R_{ON}	$(L/W) \times [1/(\mu_n Q_{ON})]$	1
τ_d	$C_g R_{ON} = L^2 C_{OX}[1/(\mu_n Q_{ON})]$	β/α^2
Maximum toggle rate	$f_{max} = 1/\tau_d$	(α^2 /β)
I_{DSS}	$(\mu_n C_{OX}/2)(W/L)(V_{GS} V_{th})^2$	$1/\beta$
J_{DSS}	$I_{DSS}/(L \times W)$	α^2 /β
E_{gate}	$(1/2)C_g(V_{DD})^2$	$1/(\alpha^2\beta)$
P_{gate}	$(V_{DD})^2/R_{ON}$	$1/\beta^2$

	$+ f_{max} \times E_{gate}$	
P□	P_{gate}/A_{gate}	$(\alpha/\beta)_2$
Figure of Merit	$P_{gate} \times \tau_d$	$1/(\beta\alpha_2)$

7.7.7.3. Scaling factors for all the three scaling models.

In this section we will tabulate the scaling factors for all the three scaling models namely:

1. Constant Electric Field Model also known as Full Scaling Model. Here $\alpha = \beta$.
2. Constant Voltage Model also known as Fixed Voltage Model. Here $\beta = 1$.
3. General Scaling Model. Here two scaling factors are kept different. This is most realistic and generally accepted by Industries.

Table 7.7.7.3.1. Scaling Factors for long channel devices for the three scaling models.

Parameter	Full Scaling	General Scaling	Fixed Scaling
W,L,D	$1/\alpha$	$1/\alpha$ for W&L, $1/\beta$ for D	$1/\alpha$
V_{DD} , V_{th}	$1/\alpha$	$1/\beta$	1

N_B*	V/d^2	α	α^2/β	α^2
Area/ device	WL	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha^2$
$C_{ox} = C_{\square g}$	$1/D$	α	β	1
C_g	$C_{ox}WL$	$1/\alpha$	β/α^2	$1/\alpha^2$
k_n or k_p	$C_{ox}W/L$	α	β	1
I_{DSS}	kV^2	$1/\alpha$	$1/\beta$	1
t_d	C_gV/I_{DSS}	$1/\alpha$	β/α^2	$1/\alpha^2$
P_{gate}	C_gV^2/t_d	$1/\alpha^2$	$1/\beta^2$	1
E_{gate}	C_gV^2	$1\ \alpha^3$	$1/(\beta\ \alpha^2)$	$1/\alpha^2$

*From Eq.7.7.7.1.1 :

$$N_B = \frac{2\epsilon_0\epsilon_{Si}(\varphi_{B0} + V_{DD})}{q\,d^2}$$

SSPD_Chapter 7_Part 7_Scaling of MOS circuits_concluded.

SSPD_Chapter 7_Part 7_Scaling of MOS circuits_concluded. gives the definition of half-pitch and node. It gives limitation posed by Interconnect delays and the ways to overcome them. It gives importance of thermal management because of exponentially growing dissipation per die with scaling. It also gives the new challenges being posed due to scaling especially the sub-threshold leakages and the use of High K-metal to overcome it.

SSPD_Chapter 7_Part 7_Scaling of MOS circuits_concluded.

7.7.8. Definition of half pitch and node.

International Technology Roadmap for Semiconductors uses ‘node’ to refer to the smallest feature size on Logic Chips. This is the length of the Gate of MOSFET. Whereas in Memory Chips ‘half pitch’ is used to define the smallest feature size.

Half-pitch is the half distance between two adjacent aluminium pathways. Nodes and half-pitch are illustrated in Figure 7.7.8.1

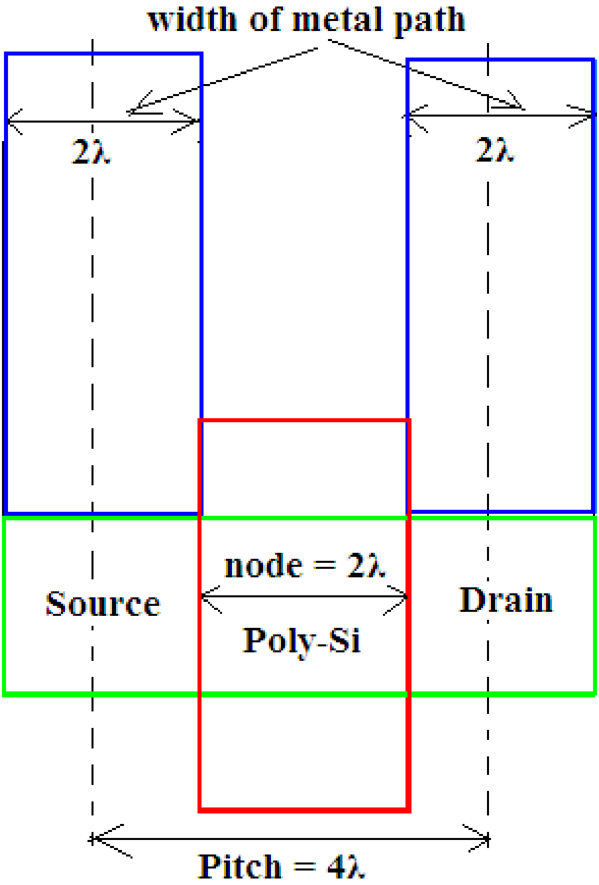


Figure 7.7.8.1 Illustration of Pitch and Node.

As is evident from the figure, Half-Pitch = Node. But with the evolution of Technology the MOS structure has been modified and today Half-Pitch > Node. In Figure 7.7.8.2 we see the evolution of MOS structure from 1970 to 2005.

As we see that with scaling LDD(Lightly Doped Drain) had to be introduced under the edges of the

Gate to mitigate the short channel effect. With the scaling of the dimensions, V_{DD} was not proportionately scaled down under General Scaling Scheme as a result sharp electric field is created which results in “hot electrons effect”. Hot Electron Effects has deleterious effects on the Device. These are the following:

1. It injects electrons into the oxide layer leading to instability of the device.
2. It limits the velocity to scatter limited velocity which in effect saturates the electron velocity and we forfeit the advantage of reduction in transit delay with scaling. In long channel devices transit delay scales down by $1/\alpha$ in constant field scaling and by $1/\alpha^2$ in constant voltage scaling.
3. In short channel when electron velocity saturates , we can obtain the scale down by $1/\alpha$ in constant field scaling but we cannot obtain $1/\alpha^2$ scaling in constant voltage scaling.
4. Also under short channel effect channel length modulation parameter gets aggravated.

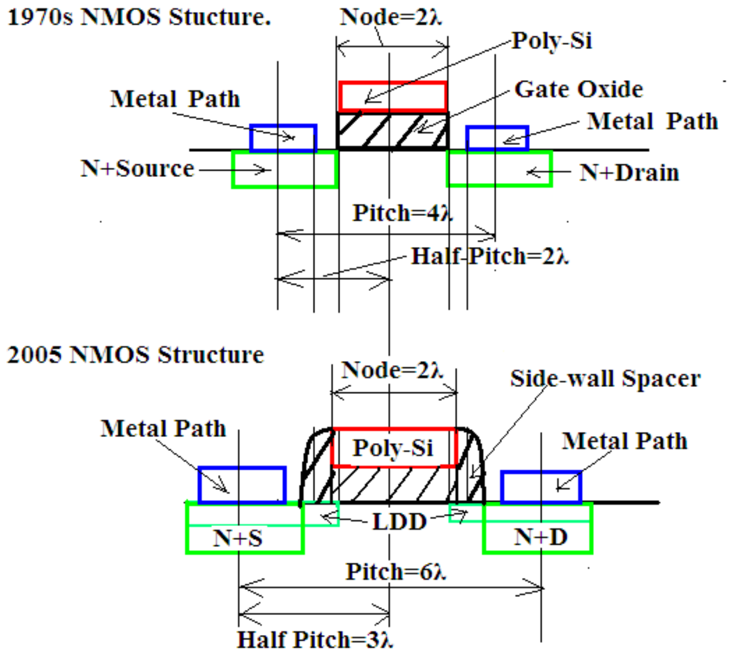


Figure 7.7.8.2. Evolution of NMOS from 1970s to 2005.

To mitigate the short channel effects we introduce Side-Wall Spacer at the edges of the Poly-Si Gate and LDD(Lightly Doped Drain) on source and drain side as shown in the lower half of Figure 7.7.8.2.. LDD extends below the the edge of the Poly-Si Gate. This has been achieved by tilted angle implant.

By grading the doping of Drain and Source from N+ to N- (LDD), we allow a more gradual drop of voltage leading to the reduction in the peak of the electric field along the channel. Even modest reduction in field leads to significant improvement in the performance of the device. It suppresses the hot electron generation, it prevents the saturation of the drift velocity of the electrons in the channel and the shallow junction of LDD helps minimize the chanel length modulation parameter.

The net result of this modification has been that in recent years Half-Pitch > Node as seen in the following tables.

In Table 7.7.8.1 we give the scaled down dimensions of MOSFET over the last 4 decades.

Table 7.7.8.1.Dimension Scaling in MOSFET over the last 4 decades.

Year	1967	1997	1999	2001	2003	2006
L(µm)	10	0.25	0.18	0.13	0.1	0.07
DRAM(Gbit/cm2)	64M	0.18	0.38	0.42	0.91	1.85
Jn.Depth(xj nm)	1000	100	70	60	52	40
Interconnection Pitch(nm)	2000	600	500	350	245	130

In Table 7.7.8.2. we give the scaling in terms of Half-Pitch and Node.

Table 7.7.8.2. Evolution of Nodes and Half-Pitch

with advancing Generations.

Year	Node(nm)	Half-Pitch(nm)	Gate Length*(nm)
2009[a]	32	52	29
2007[a]	45	68	?
2005[b]	65	90	32
2004[b]	90	90	37
2003[b]	100	100	45
2001[c]	130	150	65
1999[c]	180	230	140
1997[d]	250	250	200
1995[d]	350	350	350
1992[d]	500	500	500

a-ITRS data 2008 update.

a-ITRS data 2006

a-ITRS data 2001

a-ITRS data 1997

*The physical length has become smaller than the printed length.

7.7.9.Limits due to Interconnection.

In Figure 7.7.9.1 we have defined the width(W), thickness(t) and spacing (S) of the metal pathway/ metal interconnection at the same level and height (h)of the oxide layer separating two adjacent level metal pathways.

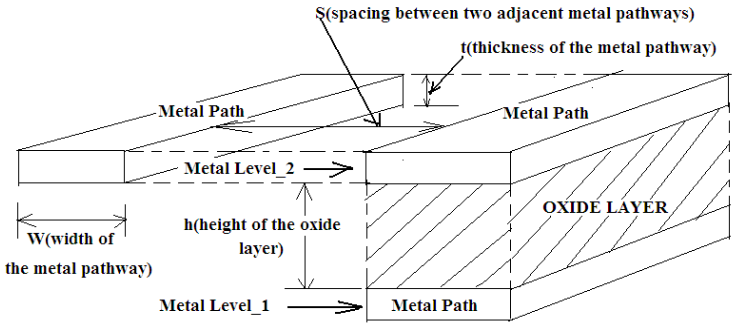

S(spacing between two adjacent metal pathways)
t(thickness of the metal pathway)
Metal Path
Metal Path
Metal Level_2
OXIDE LAYER
W(width of the metal pathway)
h(height of the oxide layer)
Metal Level_1
Metal Path

Figure 7.7.9.1. Definition of the Interconnection Dimensional Parameters for parallel interconnections at the same level and parallel interconnection at two adjacent levels.

We have two scaling models. One model is fixed thiockness and the other is scaled thickness.

Wire length will be short for 'local innterconects' and will be chip long for 'global interconnects'.

Die size will be scaled by Dc = 1.1. As we scale down, the die size is scaled up so as to build more and more complex 'System-on-Chip'.

In Table 7.7.9.1 we give the performance of local and global interconnects.

Table 7.7.9.1. Influence of Scaling on Interconnect Characteristics.

Parameter	________	Var.'t'	Const.'t'
Width(W)		1/S	1/S
Spacing(S)		1/S	1/S
Thickness(t)		1/S	1
Interlayer Oxide height(h)		1/S	1/S
Characteristics per unit length			
Wire Resistance per unit length	$R_W = 1/(Wt)$	S_2	S
Fringing Cap.per unit length	$C_{Wf} = t/S$	1	S
Parallel Plate Cap.per unit length	$C_{Wp} = W/h$	1	1
Total Wire Cap per unit length	$C_W = C_{Wf} + C_{Wp}$	1	1 to S
Unrepeated RC per unit length	$t_{WU} = R_W C_W$	S_2	S to S_2
Repeated RC per unit length*	$t_{WR} = \sqrt{(RC\ R_W C_W)}$	$\sqrt{S}$	1 to $\sqrt{S}$
Cross Talk	t/S	1	S

(Noise)			
Local/scaled Interconnect Delay			
Length(L)	L	1/S	1/S
Unrepeated Wire RC delay	$L^2 t_{WU}$	1	1/S to 1
Repeated Wire RC delay	$L t_{WR}$	$1/\sqrt{S}$	$1/S$ to $1/\sqrt{S}$
Global Interconnect Delay			
Length(Chip length)	L	D_C	D_C
Unrepeated Wire RC delay	$L^2 t_{WU}$	$S^2 D_C^2$	$S D_C^2$ to $S^2 D_C^2$
Repeated Wire RC delay	$L t_{WR}$	$D_C\sqrt{S}$	D_C to $D_C\sqrt{S}$

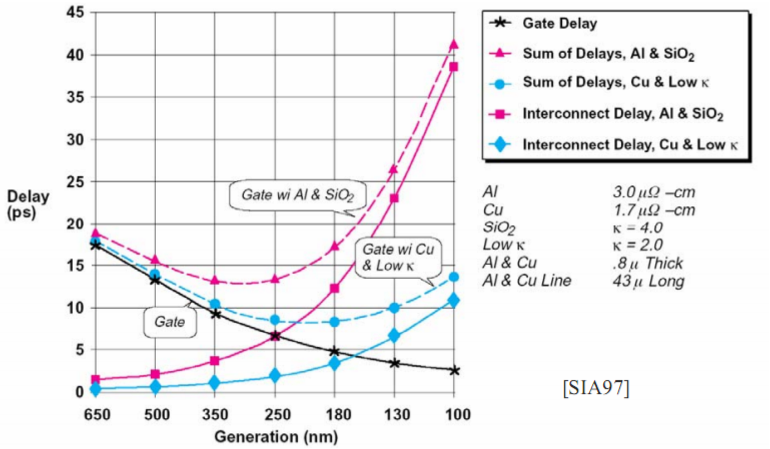
*Asuming constant field scaling of gates.

7.7.9.1.Interconnect Delays are limiting the performance with advanced Generations of Technology

Inspection of Table 7.7.9.1 leads us to the following conclusions

1. Wire Capacitances per micron is remaining constant at 0.2femto Farads/μm. This is roughly 1/10 of gate capacitance;
2. Local wires are getting faster but not quite as fast as the Gates but that is not a major problem.
3. Global wires are getting slower.

In Figure 7.7.9.1 we make a comparative study of the Gate Delays, Interconnect delays and the overall delay. We clearly see a minimum dealy of 12 psec at Node 250nm for Al and SiO_2 and a minimum delay of 8 psec at Node 180nm for Cu and low-K interconnect. This is the reason why IBM has completely switched over to Copper and low-K interconnects.



[SIA97]

Figure 7.7.9.1. Comparative study of Gate Delays and Interconect Delays.

7.7.9.2.Ways and means for overcoming the Interconnect Bottleneck of cost and

performance.

Some of the logical solutions for overcoming the Interconnect bottleneck are:

1. More levels of interconnect. We had to resort to 7to 8 levels of metallization at 100nm Node.
2. Hierarchy of design rules for local, semi-local and global routing.
3. Once the practical limits have been achieved we can further improve by going for Copper interconnects and low refractory index dielectric materials for higher propogation velocity. These dielectrics could be organic or aerogols.

7.7.10. Thermal Management due to excessive heating consequent to scaling.

Scaling reduces power dissipation per Gate but scaling increases the number of gates per chip. This coupled with higher speed scales up Chip Power Dissipation. Therefore Thermal Management becomes the central theme of Advanced Electronic Systyems.

In 1960, we had SSI chips which dissipated 0.1W to 0.3W per chip. These were cooled by Air and Liquid Cooling Techniques.

In 1980 we had BJT LSI and CMOS VLSI chips which dissipated 1 to 5W.

In 1990 CMOS VLSI were dissipating 15 to 30 W of heat.

CMOS µP chips at 1 to 2 W level were managed by heat sinks and heart spreaders.

When the dissipation level reached 2 to 3 W/cm^2 totalling to 300W at substrate level in multi-chip modules there we went for Liquid cooling.

Water cooled multi-chip modules evolved to refined indirect liquid cooling designs.

In 1990, third generation cooling modules were introduced which could handle 2 to 5kW dissipation in 225 to 1000cm^2 footprint.

7.7.10.1.Methods of Cooliing .

Some of the popular methods of cooling which were adopted were:

1. Compact heat exchangers;
2. Miniature refrigerators in Si – cooling system was imntegrated with the chip.
3. Direct liquid cooling of bare chips.
4. Chip packages in dielectric liquid;
5. In late 1980s, low cost , air-cooled multi-chip modules became popular. These had spatial and volumetric cooling facilities.
6. Water cooled modules were also developed.

The method of cooling was categorized according to the product categories.

- Low cost electronic systems costing less than $300 had bouyance, induced natural circulation of air method of cooling.
- Handheld equipments costing less than $1000 had heat spreaders. This would suffice for 1 to 2 W ICs.
- Desktop and note-books computers costing less than $3000 had several options.
    - One option was heat sink with air-cooling.
    - Second option was cooling by remotely located fans.
    - Third option was Fan cooled heat sinks.
    - Fourth was natural convecting air, circulating past low-fin heat sinks, heat pipes and metal cases. This sufficed for 3 to 5W per chip dissipation.
- High performance Systems costing more than $3000 had forced cooling systems.
- The harsh environment had still more elaborate methods of cooling.

By 1997, IC power dissipation reached 100 to 150 W heatdissipation. Chip back side is allocated to heat removal. The heat removal may be done by any one of the following methods:

- Conducting away the heat through solids with

high thermal conductivity;
- Through convection the heat is removed;
- Through pumped liquid transport.
- Throuugh Vapour diffusion.

Thermal and Thermofluid modeling software is being accepted for simulation based development and implementation of advanced packages. High performance, air-cooled heat sinks will dominate future applications.

7.7.11.The Challenges being posed by Scaling.

7.7.11.1. High-K dielectric for Gate Insulation.

The biggest challenge is the unacceptably thin gate oxide with advanced scaling.

Table 7.7.11.1 gives the thicknesses of gate oxide with advanced Generation of MOS devices.

Table 7.7.11.1. The Gate Oxide Thickness (D nm) in advanced Generations of Devices.

Generation	5μm	2μm	1.2μm	90nm	45nm	30nm
D(nm)	88nm	40nm	22nm	1.3nm	?	?

We see that as we reach 90nm Generattion of Devices, the oxide layer becomes:

$$1.3nm = 13\mathring{A} = \frac{13\mathring{A}}{2.6\mathring{A}} = 5\ atomic\ layers.$$

At 5 atomic layer thick Gate Oxide direct Quantum Mechanical tunneling is enabled leading to high Gate leakage and increased static power dissipation. The only solution was to make it physically thick but keep it electrically thin so that threshold voltage is unaffected., So we go for high K dielectric material for gate insulation.

Since Gate Capacitance has to remain constant therefore:

$$\frac{\epsilon_{OX}}{D_{OX}} = \frac{\epsilon_{new}}{D_{new}}$$

Therefore:

$$D_{new} = D_{OX}\frac{\epsilon_{new}}{\epsilon_{OX}}$$

So if a high K material $5\times$ the dielectric constant of SiO_2 is used we will have $5\times$ thick oxide layer that is 6.5nm thick dielectric material at the Gate which will completely cutoff the leakage and hence we normalize the statric power dissipation to nW in CMOS logic..

High K dielectrics are HfO_2 with K=15 to 30 or $HfSiO_2$ with K= 12 to 16..

7.7.11.2. Poly-Silicon Gate should be replaced by Metal as Silicon Dioxide replaced by High-K insulator.

The scaling forces SiO_2 to be replaced by high-K insulator but it simultaneously demands Poly-Si to be replaced by Metal. As we progress with scaling Poly-Si becomes less effective as Gate. There are three reasons for this:

1. Because of Fermi-level pinning we get higher threshold voltage;
2. High-K material has high elasticity hence results in higher phonon or lattrice scattering. This results in the deterioration of channel conductivity;
3. Poly-Si has poor bonding with high-K insulator.

Therefore Poly-Si has to be replaced by Metal again as it was in the start of IC

Technology.

7.7.11.3.Strained Silicon for enhanced channel mobility.

To further solve the problem of reduced channel mobility, Uniaxial Process induced Stress could be used for enhanced mobility and hence for higher channel conductivity.

7.7.12. Concluding Remarks on Scaling.

Table 7.7.12.1. gives the typical scaling factors which have been used for the advanced Generations of the Devices.(Reference:”Device Scaling:The Treadmill that fuelled Semiconductor Industry Growth” by Pallab Chatterjee, i2 Technologies, Inc.)

Table. 7.7.12.1.Semiconductor Association Association 1992 Overall Road Map Technology Characteristics.

	1992	1995	1998	2001	2004	2007
Node(2λ) nm	500	350	250	180	120	100
Scaling factor(α)	-	1.43	2	2.8	4.16	5
Desktop, V_{DD}(V)	5	3.3	2.2	2.2	1.5	1.5
Portable, V_{DD}(V)	3.3	2.2	2.2	1.5	1.5	1.5
Desktop, (β)	-	1.515	2.27	2.27	3.33	3.33
Portable,	-	1.5	1.5	2.2	2.2	2.2

(β)						
High Per.-P_D(W per die)	10	15	30	40	40-120	40-200
Portable-P_D(W per die)	3	4	4	4	4	4
Clock(MHz)-off chip	60	100	175	250	350	500
Clock(MHz)-on chip	120	200	350	500	700	1000
f_0 is scaled by α^2/β-offchip	60	81	105	207	312	450
f_0 according to SIA	60	100	175	250	350	500
f_0 is scaled by α^2/β-on-chip	120	162	211	414	623	901
f_0 according to SIA	120	200	350	500	700	1000

Power per die(W)	10	15	30	40	80	120
No. of Gates per die(SIA)	300k	500k	2M	5M	10M	20M
P_g(power per gate)$\times 10^{-5}$W	3.33	3	1.5	0.8	0.8	0.6

In addition to speed and packing density, scaled CMOS provides new opportunities in low power management. Speed can be traded off fiir reduction in power consumption.

40N. AA NiMH Battery Cells are required for FULL MOTION VIDEO PLAYER in 0.5μm technology.

1No. AA NiMH Battery Cells are required for FULL MOTION VIDEO PLAYER in 0.1μm technology.

This gives the kind of accessibility which can be developed for the electreonic goods with Scaling.

The Integrated Circuit Technology of Future Generations-Part 1.
This Chapter describes the deep impact Si-IC Technology has had on Computer Technology, Communication and Entertainment technology. But the further miniaturization of Si -IC Technology will hit a wall. If the post-Industrial pace of economic growth has to be maintained then Si-IC Technology will have to be replaced by Graphene IC Technolgy which appears to be one of the possible contender as an agent of main driver of Human Society and its economy.

Chapter 8. The IC Technology of Future Generations.

Once the Si IC Technolgy hits the wall by about 2025, new materials and new architecture will have to be invented and used so as to continue the miniaturization as envisioned by Gordon Moore of Intel's fame.

8.1. **Project Proposal on Graphene Transistors: fabrication, testing , characterization and its integration into future generation Integrated Circuit Chip.**

1.1.Project Proposal.

Electronics Devices, Circuits and Systems and their related services are worth $5 Trillion which is 10%

of total World’s real money market worth $50 Trillion hence semiconductor industry remains the main engine of growth of the World Economy. Electronics Circuits and Systems evolution and growth are driven by constantly shrinking device geometries and increased functionality that larger and denser Silicon IC Chips provide. This is becoming increasingly challenging and prohibitively costly and scaling below 9nm will come to a dead end sometime in near foreseeable future perhaps by 2025AD because of the physical limit. Graphene IC Chips are one of the alternatives which may become the driving force for further evolution of Electronics Systems and keeping Electronics Industry in its commanding position in World Economy and therefore possibilities have to be explored for developing Graphene IC chip at the earliest.

1.2.A brief history of Electronics Engineering and its part played in ushering Third Information Revolution*.

[*First Information Revolution was the invention of Phoenician alphabets in BC700. This laid the foundation of all European languages. The Second Information revolution was the invention of the rotating press by Gutenberg a German Mechanic, in 1450. Hence this is also known as Gutenberg Revolution. Before the invention of printing press, reading and writing was confined to a select group of people but printing press gave access to all kinds of information to a much larger section of population.]

In 1895 the discovery of electron during the study of Cathode Rays by J.J.Thomson laid down the foundation of a new discipline of Engineering namely Electronics Engineering where we dealing with electron conduction in vacuum and solids. This opened the possibility of electronic information processing. Till 1895 information processing was confined to brains and in a limited way through mechanical calculators and through Morse code and Telegraph.

1.2.1.Electronics Engineering born as a discipline in 1903 with the invention of Vacuum Tube Diode.

In 1903, John Ambrose Fleming invented Vacuum Tube Diode. This was followed by the invention of Triode by Lee de-Forest in 1905. The first RC-coupled Amplifier was invented soon after which became the wherewithal for long distance telephony and Public Service Telephone Network transformed into a Public Service Utility. In 1911, the first long distance telephone call was made from New York to Denver, Colorado, using loaded cables. Heaviside “distortion-less transmission condition” was used for constructing loaded cables. A telephonic call made on loaded cables did not suffer from phase distortion hence it was clearly audible even after

2000 miles of propagation without amplification. Subsequently RC-coupled Amplifiers were used in repeater stations along with loaded cables. In 1913 this made a long distance telephone call possible from New York to San Francisco over a distance of 4000 miles. This was the dawn of Tele-communication Era.

In 1945 first Electronic Numerical Integrator and Calculator (ENIAC) was built and demonstrated. This used 18000 Vacuum Tubes, weighed 27 Metric Tonnes, consumed 200kW and occupied a large hall.

In 1952 UNIVAC I , first main-frame computer based on vacuum tubes, was introduced for office applications. This weighed 7.26 Metric Tonnes, used 5000 valves and carried out 1000 calculations per second.

In Vacuum Tube era, Electronics Systems remained confined to Telecommunication, Radio and TV broadcast. Computers remained too expensive and cumbersome for common man’s application. It was used only for niche applications.

1.2.2.The dawn of Solid State Era.

With the invention of Germanium Transistor in 1947, the Solid State Era was ushered in. This helped proliferate the application of Communication and Computer Electronic systems by leaps and bounds because of volume compactness and much lower power consumption.

In 1959, the invention of Silicon Integrated Circuit Chip by Robert Noyce of Fairchild and Jack Kilby of Texas Instrument greatly accelerated the pace of proliferation of Electronic Systems and widened the extent of proliferation to different fields of human life. Today Electronics Discipline is so widely pervasive that no part of human society remains untouched.

Solid State Era, particularly the IC Technology, became the harbinger of third wave of civilization† and provided the wherewithal for the Third Information Revolution.

[†Agriculture brought the first wave of civilization and Industry brought the second wave of civilization.]

1.2.3.The growth of Solid State Electronics Systems.

In 1956, Sperry Rand,USA, introduced transistorized UNIVAC II. This was the second generation computer but it was IBM1401 main frame computer in 1959 which brought a spike in computer applications for business offices. In 1964 this was completely supplanted by System 360. It consolidated software, peripherals and support in on one compatible and scalable family of computers.

In 1960, Digital Equipment Corporation (DEC) introduced the mini-computer PDP-1 at a much lower cost for office applications. This was of cabinet size. In 1964 the Table-Top version PDP-8 was introduced. This was 8-bit computer. Here data word and address word, both were 8-bit wide. In 1970 PDP-11 with 16-bit address word was introduced. This has a much large memory capacity.

In 1971, the first µP chip 4004 with 4-bit data and 8-bit address word was introduced..In 1977 with the invention of Personal Computer named APPLE-II based on 8080 INTEL µP chip brought a sudden spurt in home applications of computer. This had 4kB RAM and 1MHz clock speed.

With the growth of IC Technology from one generation to another, the electronic circuits speed acceleration and microminiaturization pace was so stupendous that today a Personal Computer has surpassed Super-computer capability as seen in Figure 1.

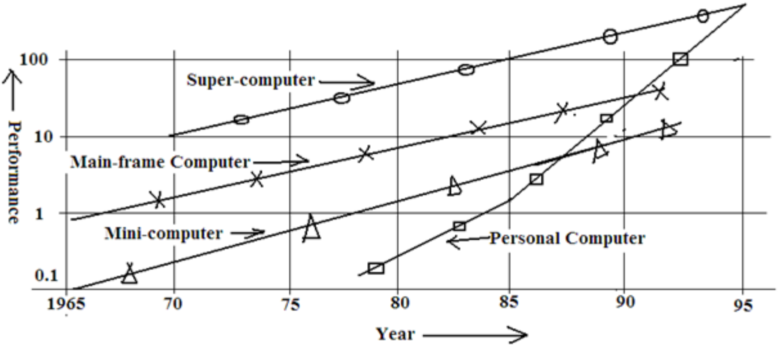


Figure 1. Exponential Growth of Personal Computer due to the rapid development of I.C.Technology.

In 1995, personal computers surpassed the performance of super-computers. Since in Desk Top Computer, which is a personal computer is, physical space and power source is at premium hence the growth in I.C. Technology and its concomitant advantage in terms of exponentially increasing logic density and memory capacity, reduction in propagation delay and reduced power consumption had the largest impact on Personal Computers as is evident from Figure 1. Chunky, nerdy, power guzzling cumbersome Computers were transformed into stylish, high tech power smart gadgets. This was mainly due to the microminiaturization, speedier computation and portability feature of I.C. Technology which has been advancing according to Moore's Law.

1.2.4.Moore's First Law and Second Law.

The market forces compelled I.C. manufacturers(such as Bell Labs, IBM, Taiwan Semiconducting Manufacturing Corporations, Intel and SAMSUNG) to invest heavily in improving the lithography techniques, design automation, device fabrication technology moving from wet to dry processing and in automation of bonding and packaging. This resulted in doubling of logic density, improvement of circuit performance by 40% and quadrupling the memory capacity every 2 Years. I.C.Technology seemed to follow this trend for the first two decades hence Gordon Moore of INTEL formulated it in a Law which now is called Moore’s First Law [2 Plummer J., Griffin P. “Material & Process Limit in Silicon VLSI Technology”,Proceedings of the IEEE, **89**, No.3, 240-258, March 2001].

As sophistication and complexity of I.C. Circuits has increased so has the cost of Fabrication . In 1986 the fabrication cost of μP 80386 containing 250,000 transistors was $200million. In 1996 the fabrication cost of Pentium I chip containing 6 million transistors has soared up to $ 2 billion. This is known as Moore’s Second Law.

1.2.4.1.The impact of Moore’s Law on World Economy.

Today Semiconductor Industry has become the main engine of growth of World Economy. The total World Economy is worth $ 50 trillion in which Semiconductor Industry and all its related services are $5 trillion worth, almost 10% of World’s GDP. This has been made possible because the market forces have compelled I.C. manufacturers to follow Moore’s Law and we can afford no slackening in this pace of growth if we are to maintain Electronics Industry as the main driver of World Economy.

1.2.4.2.CMOS Technology- the workhorse of Semiconductor Industry for last 40 years.

CMOS Technology because of nano-watt-logic characteristics and its inherent higher density of integration (because no isolation boundary diffusion is required) as compared to those of Bipolar Junction Transistor (BJT) has become the workhorse of Semiconductor Industry for last 40 years.

Memory Chip Capacity has exponentially increased. SRAM uses 6-T(transistors) to store a bit of information. SRAM is the dominant form of embedded memory. It occupies 60-70% chip area and constitutes 75-85% transistor count. Its capacity must scale up at an equivalent rate as the logic gate density is being scaled up.

From 1960 to the present day, transistors per chip has increased from few transistors to billion transistors functioning on a single monolithic chip completing the task of a super computer.

Memory size has increased from core memory of 4kB in PDP-8(minicomputer) in 1965 to 3.21GB RAM in DELL Opilex with 2 Quad CPU Q8400@2.66GHz in 2010. Disk Memory has increased from 32-128kB disk in PDP-8e Computer in 1972 to 10MB disk in IBM XTPC in 1982. In 2001 PC had 40GB hard disk, in 2003 it was 120GB hard disk and in 2010 it was 400GB hard disk.

Table 1. gives the trend of improvement in circuit complexity and performance over the years..

Table 1. The growth of CMOS IC Chips from 1997 to 2012 according to ITRS_2003.

	1997	99	01	03	06	09	2012
Feature size(µm)	0.25	0.18	0.15	0.13	0.1	0.07	0.05
VDD(V)	1.8 2.5	1.5 1.8	1.2 1.5	1.2 1.5	0.9 1.2	0.6 0.9	0.5 0.6
Transistors/ Chip(M)	11	21	40	76	200	520	1400
DRAM(b/ Chip)	167M	1.07G	1.7G	4.29G	17.2G	68.7G	275G
Die Size(mm)	300	340	385	430	520	620	750
Global Clock(Hz)	750M	1.2G	1.4G	1.6G	2.0G	2.5G	3G
Local Clock(Hz)	750M	1.25G	1.5G	2.1G	3.5G	6G	10G
Max Power/ Chip(W)	70	90	110	130	160	170	175

1.2.5.The impact of micro/nano miniaturization on Electronic Systems.

Due to Silicon Transistor scaling both vertically as well laterally, it has been possible to give greater functionality to IC Chips. This has resulted into more powerful computers both in terms of memory space as well as in computational speed in a more compact volume.

RF and HF application data rates and frequencies continue to increase for in each new technology generation.

In 70’s the growth of personal computer fuelled the CMOS IC Chips based on CMOS scaling.

In 21st century, Internet, high speed communication, the convergence of data, video, audio and mobile communication are fuelling CMOS scaling and has brought in focus the need for cheap, high speed and low power devices. SiGe devices are filling this need.

In 1971, four LSI chips built a complete functional computer. A typical 4004 LSI chip was based on 10μm PMOS(E) Technology Process. It had 2300 transistors and carried out 92.000 Instruction per second.

In 1975, the functionality of all the four IC Chips was put in one chip μP 8086 chip. This brought a big leap in microminiaturization, compactness and portability of electronics systems.

In 1980, μP80286 and μP80386 had ×87 Floating-point hardware on a separate chip.

In 1989, tightly pipelined 8086 architecture was introduced as μP80486 and this included ×87 Floating-point hardware.

In 1990, single instruction multiple data started as a separate chip. SUN Microsystem moved it on-chip when VIS was introduced in ultra sparc.

In 1992-93, Pentium had CPU + Memory + I/P & O/P port built on the same chip plus it had a second ALU. It was INTEL first super scalar processor.

L1 Cache was moved on the die in Pentium 2 architecture.

Latest Pentium Architecture started with 400MHz Bus and 256kB L2 Cache later increased to 800MHz and 2MB Cache. This model contained 42 million transistors.It used 0.18μm Process and came in 423 pin and 478 pin PGA packages.

INTEL's first Pentium 4 chipset was 850 and supported Rambus Memory (RDRAM) but subsequent chipset switched to DDR SDRAM. Eventually Floating-point hardware moved on the die.

Thus new features were added in scaled down chip to add performance-enhancing functionality to a processor die.

Processor’s Power consumption and dissipation has been on the increase along with die size as the logic density and memory density has been steadily increasing. This is creating the limitation of Fire Wall- maximum limit of dissipation per Si chip.

Using multi core processors along with threading software, this problem of Fire Wall limitation has been circumvented. Multi-core processor is a combination of smaller, cheaper and less power hungry processors working in parallel to do the same job.

Today Mobile Computing demands that more functionality be added to a single die not for increased performance but for ideal mobile computing environment. This will mean combining non-volatile memory, volatile memory, CPU and multiple types of wireless capabilities (blue tooth, 802.11b, 802.11g, GSM e.t.c) on a single die.

Improved circuit performance and more complex functionality with reduced cost has enabled IC Chip to penetrate every domain of life.

With nano-miniaturization and ultra deep-submicron technology we are moving towards a new breed of complex, highly integrated systems known as Systems-on-Chip (SOC). In SOC, RF and Base-Band functions are integrated on the same chip. This gives compact and inexpensive solutions for wireless connectivity like cellular phones. SOC is a complete end product. It may interface the real world hence they integrate analog components and in future may include Opto/MEMS.

To create SOC designs with the required complexity, designers must use pre-designed Intellectual Propriety(IP) Blocks referred to as macros, cores and virtual components. This means reuse of cores. The more design reuse we have , the faster delivery time of SOC solutions.

Nanoelectronic Components such as sensors and actuators have to be incorporated in SOC. These are realized through MEMS. Thus micro/nano miniaturization has helped us achieve new level of integration among diversified systems of electronics, mechanical, optical and fluidic elements.[34,35 Murari, B. "Integrating Nanoelectronic Components into Electronic Microsystems, IEEE Micros, **23**, No.3, 2003, pp.36-44]

To meet computational intensive applications and to meet the requirement of low power, high performance systems, the number of computational resources has enormously increased and hitherto Si IC Technology has been the enabling technology. But with such high circuit density, increasingly the interconnections and data buses are becoming the major challenge. A shared bus interconnection where an arbitration logic to serialize several bus access requests is adopted to communicate with each processor. In such an environment where number of bus requesters is large and required Band Width for interconnections is more than the current bus, there scalable BW requirement can be satisfied using on-chip packet-switched micro-network of interconnects known as Network-on-Chip (NOC) architecture. The scalable , modular nature of NOC and their support for efficient on-chip communication leads to NOC-based system implementation.

New products such as Software Defined Radios are coming into the market. Here there is reuse of Silicon thereby lowering the cost. Software-driven functionality offers flexibility and upgradability.

In the past chip was a component . Today it is a complete system.

With circuit miniaturization, information processing has moved from main-frame to personal computers. From personal computers it is moving to portable computing integrated in larger products such as smart phones. This kind of computing is called ubiquitous computing, pervasive computing or ambient computing. Embedded systems are the wherewithal for this new kind of computing.

Embedded systems(ES) are the computers

incorporated in consumer products(entertainment products, home appliances, medical appliances and military appliances) to perform application specific functions. It can contain microcontroller, ASIC, ASIP ,DSPs. Here the electronics are deeply embedded and must interact with the users and the real world through sensors and actuators. These computing devices must work in real time hence they have Real-Time-Operating-System(RTOS).

ES are built using a wide variety of techniques such as software, firmware, ASIC, ASIP, General Purpose Processors, Domain Specific Processors, Field Programmable Logic Devices, Complex Programmable Logic Devices, Analog Circuits, sensors and actuators. ES finds applications in Medical Electronics(pacemakers), Personal Communication Phones (mobile phones)., Automobiles(anti-lock braking systems), Avionics(fly-by-wire flight control system), Railways (high speed train control) and home appliances. ES provide SOC and high level module solution.

1.2.6.The Challenges and difficulties beyond 20μm Process Technology.

All this advancement and convergence in computation and communication was made possible due to scaling. But this is becoming more difficult as we go below 20μm process.[6 Hennessy, J; Patterson, De; "Computer Architecture", 3rd Edition, Morgan Kaufman Publishers, Amsterdam 2003].

According to International Technology Roadmap for Semiconductors (ITRS) we have the projection as given in Table 2.

Table 2.Roadmap of CMOS Technology according to ITRS-2003

Year	**05**	**06**	**07**	**08**	**09**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
LGate (nm)	32	28	25	22	20	18	16	14	13	11	10	9
	←Bulk/ PartiallyDepleted- SOI→											
				←Fully Depleted – SOI→								
							←Multi-gate→					

SRAM uses 6-T to store a bit. SRAM is dominant form of embedded memory. It comprises 60- 70% chip area and 75% - 85% transistor count. As the logic gates density is increasing correspondingly the memory density must increase.

Statistical dopent fluctuations, variation in oxide thickness and line edge roughness increases the spread in V_{Th} and I_{ON}/I_{OFF} in nano-scale regime [34 Murari 2003]. Increased leakage and parametric variations are making scaling of 6-T SRAM memory array difficult.

Due to fundamental physical limit to scaling, the era of conventional linear scaling of transistor dimensions has ended. Power dissipation and process induced variations have become the major issues for continued scaling of Si MOSFETs in future generation.

New materials and device structure are needed to continue the scaling trend as seen in ITRS-2003 Road map given in Figure 2.

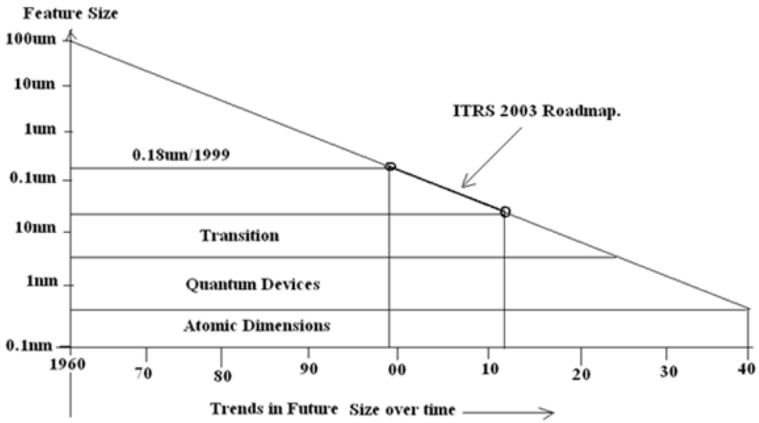


Figure 2. Trends in Future Size as illustrated in ITRS 2003 Roadmap.

According to ITRS 2003, by 2014 $L_{Gate} = 35nm$, $V_{DD} = 0.4V$ and clock frequency = 30GHz.

Shrinking geometries, low power voltages and high frequency have a negative impact on reliability. These enhanced features have increased the failure rates.

1.2.7.New methods to meet the challenges of nano miniaturization.

Multiple core with threading has been one innovation for solving the thermal management at nano level.

Copper interconnect provides higher electromagnetic threshold than Aluminum. So Industry is adopting Cu interconnects.

Fault avoidance and fault tolerance are the possible ways of improved reliability.

Fault avoidance relies on improved materials, improved manufacturing processes and improved circuit design. Cu interconnects is an example of improved materials.

SOI is the process solution for lower circuit sensitivity to particle induced transients.

Fault tolerance is implemented at circuit and system level. It relies on error detection and error correction code.

1.2.8.Future of Integrated Circuits.

Human appetite for computation has grown faster than the processors power of real life computers. We need more powerful processors just to keep up with modern applications like interactive multi-media, mobile computing and wireless communication. In addition battery power is at premium in mobile computing environment hence power aware components/systems are indispensible.

Low power design can be achieved at different level of the system- basic process/device level (threshold reduction, multi-threshold deices), at circuit level (transistor sizing, logic optimization, activity driven power down, low swing logic, adiabatic switching), at architecture level (voltage scaling, parallelism,, instruction set , signal correlation) at algorithm level(complexity, concurrency, locality, regularity,data representation) and at system level. (design partitioning and power down).

We have to simultaneously look at performance-power-cost product . This will lead to new computer architecture.

As we scale down the dimensions for each new generation technology, new areas of applications are emerging.

To continue its historical growth and continue to follow Moore's Law, the Semiconductor Industry will require advances at all fronts- from front end processors and lithography to design innovations, high performance process architecture and SOC solutions.

Further down the scaling strategy we are bound to hit a road block.

The recent development in the field of Graphene Transistors show that these could be a promising candidate for taking over the mantle for continued scaling down according to Moore's Law.[Kreupl, Franz " Carbon nano-tubes finally deliver", *Nature,*.**484**, 321-322, 19th April 2012].

The present Graphene Transistor using single walled Carbon nano-tube(SWCNT) of 9nm at a supply voltage of 0.4V switches from '0' to '1'. The current ratio of ON to OFF is 10,000 times which is excellent for LOGIC applications.. This operates at low voltage hence it mitigates the demand in energy. The only difficulty is in fabricating these devices because of non-availability of SWCNT of one flavor.

In contrast according to IRTS 2011[ITRS2011 Edition , Front End Processors

Go.nature.com/qykuh1(2011)], a similar sized Si MOSFET of 9nm gate length and operating at 0.64V will arrive in the market in 2022 and we cannot anticipate the cost.

So it is ripe time that we explore for alternatives if the present rate of miniaturization and development has to be maintained.

Graphene IC Part 2_Graphene IC+Graphene Allotropes
Graphene IC Part 2 gives the possible Graphene IC architecture and the material science of Graphene and its various allotropes.

8.2.Is Graphene based I.C. ready to don the mantle for continued scaling down along Moore's Law ?

Quoted from *"Electrons in Atomically Thin Carbon Sheets Behave Like Massless Particles"* Mark Wilson,

*Phys. Today***59** (1), 21 (2006); doi: 10.1063/1.2180163

"Microelectronics engineers are paying attention. In semiconductor heterostructures used to make FET devices, for instance, it takes million-dollar epitaxy machines and exquisite care to tie up dangling surface bonds and eliminate impurities in quantum wells. The preparation minimizes the scattering of electrons against interfaces and defects to ensure the largest electron mean-free paths in the device.

But in graphene, just 1 Å thick, scientists have a material that is relatively defect free and whose electrons have a respectable mean-free path naturally, without materials manipulation

and processing. Graphene can hardly be more low tech, and yet it still exhibits high

conductivities. “It’s really counterintuitive and remains to be understood,” comments Geim, “but

the electron wavefunction appears to localize only parallel to the sheet and does not interact with the outside world, even a few angstroms away.”

Recent experiments have demonstrated the unique electronic properties of graphene thus charting a potential route to nano-electronics based on epitaxial graphene (EG). There are two possible approaches to nano-electronics [Hass et.al.(2006)]:

Approach 1- We use CNT to make gated devices and ballistic conducting wirers. To assemble them in an IC we will require to meet the following challenges:

1. The control of the properties of individual CNT (e.g. diameter and helicity);
2. Control of inherent heterojunction impedances associated with inter connection of CNT;
3. Finally the assembly of vast networks from individual CNT devices.

Approach 2-We rely on the continued scaling of Lithographic Techniques of the present day I.C. Technology. We use this for nano-patterning of graphene into graphene based IC. Graphene has excellent transport properties (very high mobility) and permits the control of elerctronic properties such as band-gap and doping.

By low temperature epitaxy, we achieve 1cm × 1cm graphene sheet on an appropriate substrate (may be SiC with C-face). This lithographically patterned into narrow ribbons or other shapes to provide the necessary confinement for devices.

Graphene based IC will be a scalable assembly of nano-patterned EG devices such as ballistic transistors and ballistic interconnects.

Graphene is a semi-metal at micron level or larger. But at less than 100nm, electron confinement opens up the band gap. This band-gap can be tuned and adjusted according to the requirement of the application. It can be sliced wider and narrower and in different patterns according to the requirement(wire, ribbons or some other component). To avoid contact resistance problem graphene sheet can be patterned into an array of thin parallel strips or wires.[Wilson (2006)]

Berger et.al.(2004) have suggested that if suitable methods were developed to support and align graphene sheets, it would be possible to combine the advantages of nano-tube-like electronic properties with high resolution lithography to achieve large scale integration of ballistic devices.

An essential difference between nano-tubes and planar graphene ribbons is the presence of dangling bonds at the edges. Normally these would be hydrogen-terminated with little influence on the valence electron properties. However edge atoms could be passivated with donor or acceptor molecules thus tuning the electronic properties without affecting the graphitic backbone.

8.3.Material Science of Graphene and its exotic properties.

In 2004, Andre Geim and Konstantin Novoselov, two chemist at the University of Manchester in UK, peeled out a layer of graphite which we call graphene. These are flat mono-layer of Carbon atoms tightly packed into 2D honey-comb lattice a living embodiment of 2D geometry in real life and taught in our text books.

Phillip Russell Wallace, a theoretical Physicist of McGill University, Montreal, Canada, predicted electronic structure of graphene in 1947. But scientists believed that such a 2D structure would be unstable.

Graphene was known to be an integral part of 3D Graphite but graphene sheet was presumed not to exist in a free state. It was largely considered to be an “academic material” and believed to be unstable in real life. The melting temperature of the film rapidly decreases with decreasing thickness and they become unstable(segregate into islands or decompose) at a thickness of dozen of atomic layers. So atomic mono-layers were only known as an integral part of larger 3D structures usually grown epitxially on top of single crystals with matching crystal lattice. Unexpectedly in 2004 free-standing graphene could be exfoliated and follow up experiments proved that the charge carriers in graphene were indeed zero rest-mass Dirac fermions[Novoselov et.al.(2005),Zhang et.al(2005)] and that 2D layer with unusual properties had been discovered. The effective mass of electron in graphene is measured to be $0.007m_e$ [Zhang et.al.(2005)]. Electrons and Holes in grapheme are charge conjugated symmetrical particles due to time-reversal invariance just as electrons and positrons are in Quantum Electro-Dynamic Systems [Ando (2009)]. A direct consequence of this charge conjugate symmetry is that electrons and holes in graphene have equal mass, equal mobility and opposite charge and hence can be represented by the same Dirac spinor function. This is quite unlike semi-conductor devices. In semi-conductor devices electrons and holes are independent entities with their own effective mass and their own mobilities. In intrinsic Si, electron has a mobility of 1450 cm^2/(V-cm) whereas holes have a mobility of 450 cm^2/(V-cm). In addition independent Schrodinger matter wave functions have to be used for describing the two entities.

Soon after, free-standing 2D atomic crystals of other compounds were also discovered such as single layer Boron Nitride and half layer $Bi_2Sr_2CaCu_2O_x$ [Novoselov et.al(2005)].

As shown in Figure 3, graphene is a basic 2D building block for graphitic materials of all other dimensionalities. It is considered the materia prima for other forms of Carbon. By creating topological defects, it can be wrapped up into 0D fullerenes, rolled into 1D carbon nanotubes(CNT) and stacked into 3D graphites. 0D fullerenes are also known as Buckeyballs and have 60 C atoms. 12 pentagon plaquettes are required in addition to hexagonal plaquettes to produce the spherical configuration shown in Figure 4. Graphene rolled into 1D CNT ,armchair or zigzag configuration, is shown in Figure 4.

Bulk highly oriented Pyroletic Graphite , the raw material for the preparation of graphene, is a high purity semi-metal with a highly anisotropic electronic structure featuring nearly compensated low density electrons and holes with very small effective mass. In graphite, the monolayers of graphene are stacked at a separation of 3A°. The monolayer sheets are held together weakly by vender Waal’s forces and hence they can easily be peeled off by mechanical ex-foliation [Zhang et.al.(2005b)]

Graphene and bilayer have simple electronic spectra. They are both zero gap semiconductors(or zero-overlap semimetals) with one type of electrons and one type of holes. Three or more layers but less than 10 have increasingly complicated spectra. Several charge carriers appear. Conduction Band and Valence Band start notably overlapping. Thus we can distinguish between single-, double- and few(3 to <10) layer graphene as three different types of 2D graphene crystals. Thicker structures should be considered as thin films of graphite.[Geim & Novoselov (2007)]

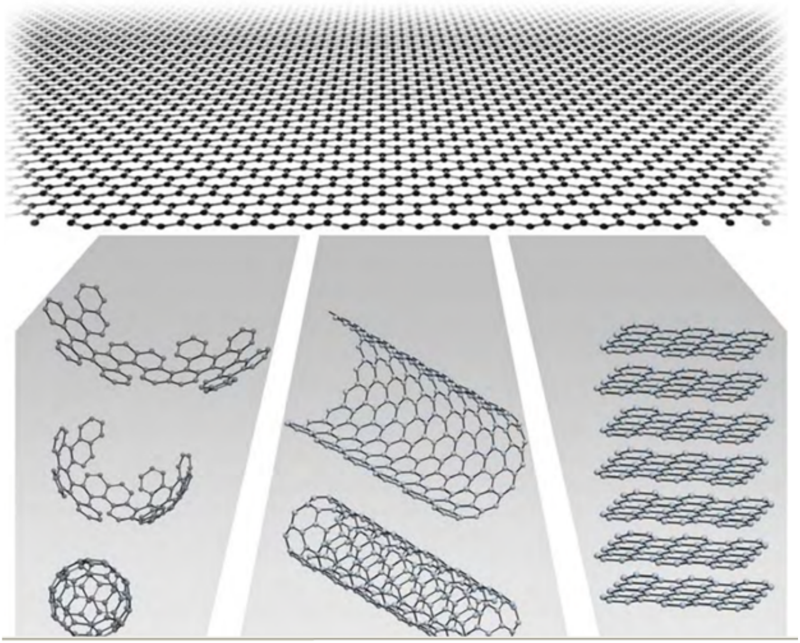


Figure 3. Graphene-mother of all graphitic forms. Graphene is a 2D building material for carbon materials of all other dimensionalities. It can be wrapped into a 0D buckeyball, 1D

nanotube or stacked into a 3D graphite. [Geim & Novoselov (2007)]

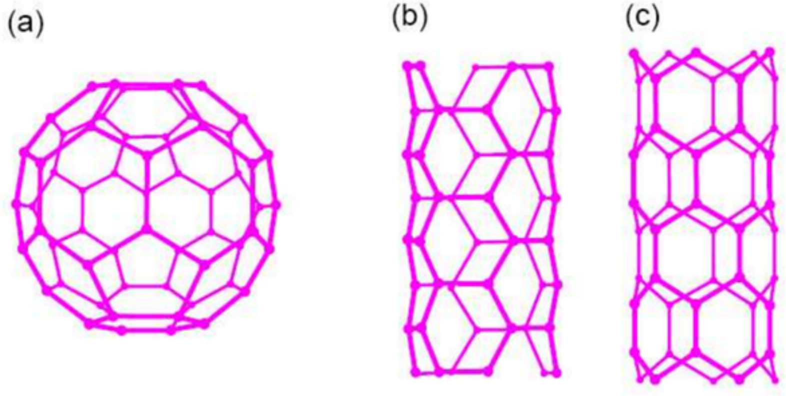


Figure 4. (a)The C60 fullerene molecule, where there are 12 pentagons necessary to produce the spherical configuration. Nano-tube of arm chair type (b) and of zigzag type(c).It is not necessary to introduce pentagons for producing nanotubes as they have zero curvature. [Pachos (2009)]

In graphite, atomic bonds between layers are much weaker than the bonds across its layer.

At 300K, mean free path of electron in graphene is several microns, an order of magnitude better than that in semi-conductors. The kind of massless electron we see in super-conductors at liquid He temperature we witness the same kind of kinematics at 300K in Graphene. Hence graphene provides a rare opportunity for doing quantum mechanical research.

It has 100× tensile strength as compared to that of Steel. It has much better thermal conductivity than that of Diamond which till now was the most thermally conductive material. It is also most closely packed structure. It is impermeable to the smallest atom like Helium but it still can stretch.

It is zero band-gap material meaning by conduction band and valence band lie close together hence graphene is a conductor but it cannot switch on and off.

In semi-conductors, because of the band- gap, photon must exceed the bandgap to cause photo-excitation and absorption hence semi-conductors are susceptible to only certain part of the spectrum. In graphene there is no such limitation. Graphene has equal absorption coefficient from UV to far IR light because it has zero bandgap.

Exploiting Graphene's optical and electron abilities along with strength and flexibility could lead to foldable plastic smartphones, cheaper solar cells or sensors that can detect single molecule of gas or identify individual DNA bases.

Silicon took decades to find its signature role in technology. Graphene's journey has just begun.

Graphene has superlative mechanical, thermal and electronic properties as shown in Table 3[Savage (2012)].

Table 3. Tensile Strength, Young's Modulus, Electron Mobility and Thermal Conductivity.

	Tensile Strength MPa		Young's Modulus GPa		Electron Mobility cm^2/(V s)		Thermal Conductivit W/(mK)
		CNT	3000				
Graphene	10^5★	Diamond	2000	Graphene	200,000	Graphene	3000 5000
CNT†	20,000						
Si	8000	Graphene	1000	GaAs	8000	Cu	300
Diamond	2500	Steel	200	Diamond	2000	Ag	280
Steel	2000	Si	180	Si	1450	Au	220
						Al	200
						Si	180
						Glass	1.8

†CNT-Carbon Nano-Tube- it can be metal or semiconductor.

★Graphene- Electrical Conductivity of Graphene is independent of 'E_F' and 'n'(carrier concentration) as long as variation in scattering length is neglected.

Therefore Graphene is metal and not zero-gap Semi-conductor.

From Table 3 , we can conclude that graphene is the thinnest, strongest, stiffest, most stretchable (almost 10%), having record thermal conductivity, permissible current density at 300K is 10^8A/cm^2 million times higher than that of copper, highest intrinsic mobility (100 times that in Silicon), conducts electricity with zero carrier density, electrons and holes carriers behave like electron and positron in Quantum ElectroDynamic Systems and also behave like Dirac quasi particles, carriers have longest mean free path of the order of micron and are most impervious hence they cannot host interstitial dopents. In effect graphene is a relativistic system goverened by Dirac Equations.

8.3.1.Graphene is a perfect conductor.[Sung & Lee,(2012)]

In conventional materials (conductors and semi-conductors) mobile carriers suffer from impurity/ defect scattering and lattice vibrational scattering also called phonon(acoustical wave packet) scattering. Mean free path(L*) in metals are 90A°,250A°,293A°,328A°,441.6A° for Li, Na, K, Cu, Ag respectively and in semiconductors L* are 729A°, 2106A°, 4645.5A° for Si, Ge and GaAs respectively.[http://cnx.org/content/m34343/]. The electron deBroglie wavelength are 5A° and 78A° in metal and semi-conductor respectively hence electron is strongly scattered by metallic lattice and weakly scattered by semi-conductor lattice [ibid]. This precisely is the reason for a much higher mobility of electron in semi-conductor as compared to that in metal. In metal mobility is only 44cm^2/(V-s) whereas it is 1450 cm^2/(V-s) in Si, 3000 cm^2/(V-s) in Ge, 8500 cm^2/(V-s) in GaAs, 77000 cm^2/(V-s) in InSb and 10×10^6 cm^2/(V-s) in quantum well heterostructures of GaAlAs. All these materials have quadratic carrier energy spectrum given by Equation 7.

In contrast Graphene has a carrier spectrum which is linear and chiral (see Section 8.4.4.2.6) with four-fold degeneracy arising from spin and valley. It has in addition a Berry Phase term arising from two sub-lattice symmetry which restricts the carriers from backscattering (see Section 8.4.1). The theoretical absence of backscattering led to the speculation that mobility could be extremely high even at 300K. This indeed was the case as proved by Bolotin et.al(2008). They measured a mobility as high as 200,000cm^2/(V-s) in suspended specimen of graphene. In high mobility graphene electrons or holes move as described in Newton’s Laws of Motion over a distance of microns: a body in motion continues to be in motion along a straight line with an uniform velocity (=c/300) unless made to act otherwise by the application of force.

The mobile carriers in 2-D honey-comb Carbon atoms array have intriguing (and conceptually novel) linear “Dirac-like” bare Kinetic Energy dispersion spectra namely [Ando (2009), derived in Section 8.4.4.2.4]:

$$E(k) = \pm\hbar v_F \left| k \right| \qquad 1$$

As we will see in Section 8.4.2, the three sp_2 hybrid orbitals of C atom in graphene constitute the σ-bond and one $2p_z$ orbital constitute the π-bond. The σ valence band is filled up but π valence band is half filled. Electrons in π-band can hop from one C site to another C site. This quantum-mechanical hopping between the symmetrical sublattices A and B (this term is defined in Sec.8.4.4.2.6) leads to the formation two cone like energy bands and their intersection near the edge of the Brillouin zone yields the conical spectra as shown in Figure 5. Because of this conical energy spectra a linear energy dispersion is exhibited described by Eq.1.

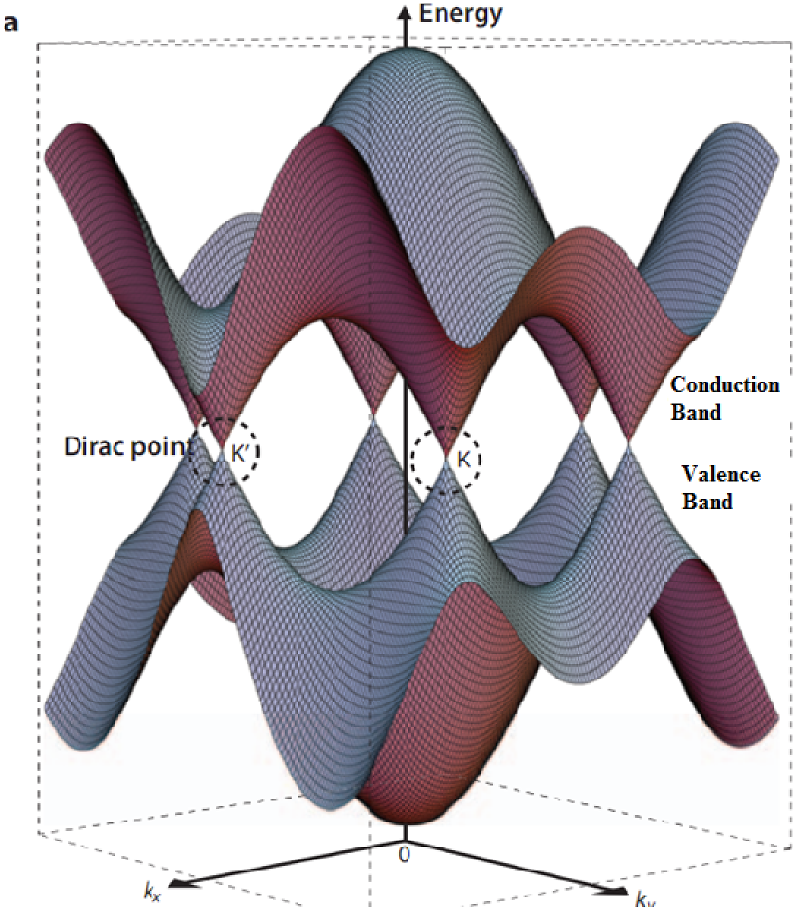
The intersection point of the two cone-like energy bands is charge neutrality point(CNP) or Dirac Point. For sublattice A the Dirac Point is ‘K’ and for sublattice B the Dirac Point is ‘K''.

Equation 1 is in contrast to the parabolic energy spectra of carriers in 3-D solids. *It is this difference in energy spectra of carriers in 2-D graphene and 3-D solids which is responsible for the unusual properties of* *graphene*.

This zero-bandgap band structure, as shown in Figure 5a and 5b, results into a linear and chiral (we will give the definition of chirality in Section 8.4.4.2.6) carrier spectrum with 4–fold degeneracy arising from spin and sub-lattice symmetry (also known as valley symmetry).

The inherent sub-lattice symmetry results into a non-zero Berry Phase term [Ando (2009)]. This non-zero Berry Phase term prevents carrier backscattering. When the direction of particle motion is rotated by 2π radians, the phase of the wave function changes by ±π which changes algebraic sign. This leads to the absence of back scattering when a particle is scattered by impurities. Backscattering corresponds to the rotation of the direction by ±π radians and the amplitudes for ±π rotation cancel each other due to the algebraic sign difference.

In CNT, the electron motion along the circumference of the tubes is quantized and electron motion is effectively one dimensional with resistance determined by back-scattering. Therefore absence of back scattering means that metallic CNT are ideal conductors with perfect conduction even in the presence scatterers.


a
Energy
Conduction Band
Dirac point
K′
K
Valence Band
0
k_x
k_y

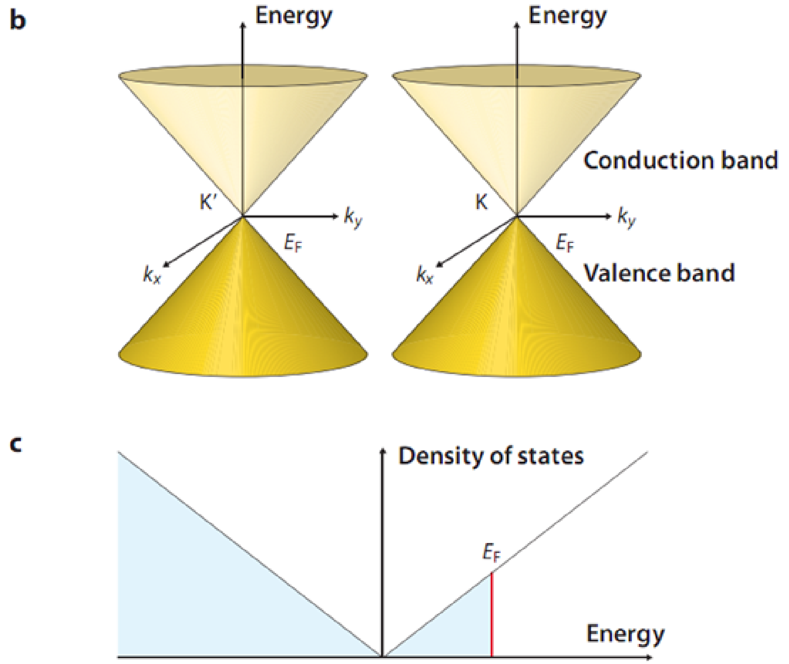


Figure 5. (a) Energy bands near the Fermi level in graphene. The conduction and valence bands cross at points K and K'. (b) Conic energy bands in the vicinity of the K and K' points. (c) Density of states near the Fermi level with Fermi energy E_F [Ando (2009)].

As a result the carriers in graphene move as massless Dirac particles , effective mass of $0.007m_e$ [Zhang et.al.(2005c)](this vanishing mass is justified in Section 4.4.2.2) at a relativistic velocity of 10^6m/s ten times faster than that in Silicon($v_{scatter_limited} = 10^5$m/s) and 1/300 of velocity of light(c) in vacuum. Hence graphene based devices consume less power and less chip area and speed up computation 3 times because of reconfigurable characteristics. The same OR gate can be transformed into EX-OR gate and again revurt back to OR gate depending on the need of the circuit [Sung & Lee (2012)].

In Silicon electrons has a mean free path of 72.9nm whereas in Graphene the mean free path is in microns. In Graphene Electron moves in a straight line across tens of thousands of atoms at ten times the speed of that in Silicon.

Metals have overlapping valence and conductive Bands, insulators have large band-gaps(several tens of eV) , semi-conductors have moderate band-gaps(~1eV) whereas graphene has zero-bandgap. This implies that it has no ON and OFF states. It is always in ON state but in laboratory we see very unusual consequences of zero-bandgap property which we will examine in a subsequent section [Sung & Lee (2012)].

There are ways by which zero bandgap can be opened and ON and OFF state can be achieved as is the requirement in Logic Gates.[Section 4.4.7]

8.3.2. Crystalline Structure of Graphene.[Chen et.al.(2009)]

Graphene is 2-D carbon ($_6C = 1s^2 2s^2 2p^2$) mono-layer built by strong carbon-carbon sp^2 bonds which provide graphene with high intrinsic strength across the 2D layer and this strength makes possible the isolation of single atomic layers. There are three isotopes of carbon 12C, 13C and 14C present in all carbon allotropes in the ratio 98.89%,1.11% and 1×10^{-10}% respectively. 12C and 13C are stable isotopes whereas 14C is radio-active isotope. 14C radio-active isotope has a half life-time of 5730 ± 40 years and they are used for radio-carbon dating of archaeological samples within 30,000 years of age. Most graphites and graphene in use today have negligible 14C. Therefore the percentage composition of stable 12C and 13C is decisive in determining the exotic properties of graphene.

In a study by Chen et.al (2012) it is found that the physics of phonons, the acoustical wave packets and the main heat carriers in graphene, have been found to be influenced by the percentage isotopic composition of the graphene. The thermal conductivity, K, of isotopically pure 12C(0.01% 13C)graphene was higher than 4000W/mK at 320K and more than a factor of two higher than the value of K in graphene sheets composed of 50:50 mixture of 12C and 13C. The experimental data is corroborated by Molecular Dynamics simulation.

In CH_4 (Methane), 2s orbital mixes with three 2p orbitals to form four sp^3 hybrids. The four sp^3 hybridized orbitals are overlapped by four hydrogen’s 1s orbital to form a tetrahedron structure with 109.5° angular spread between two adjacent apex of the tetrahedron as shown in Figure 6.

In C_2H_4 (Ethene), only two of the three available 2p orbitals form three sp_2 hybridized orbitals with one 2p orbital remaining as it is. This is triagonal hybridization. Three sp_2 hybrid orbits are coplanar and directed towards the corners of an equilateral triangle with 120° angular span. The third un-hybridized 2p orbital lies at right angle to the molecular plane of sp_2 orbits. Two sp_2 hybrid orbitals of a C atom form covalent bonds with 1s orbital electron of two H atoms. Two C atoms form π bond using 2p orbitals and σ bond using sp_2 hybrid orbital as shown in Figure 7.

Similarly three sp_2 hybrid orbitals and one 2p orbital are used to form the hexagonal benzene ring structure of C atoms in 2D plane of Graphene as shown in Figure 8. Graphene is one-atom-thick planar sheet of sp_2- bonded and 2p-bonded carbon atoms through σ-bond and π-bond respectively. These are closely-packed in a honeycomb crystal lattice. C-C bond length is 1.42A° in a single graphene layer . In a stack of graphene layers which is graphite the interplanar spacing is 3.35A°. There is zero band-gap between valence band and conduction band as shown in Figure 10. In contrast Diamond is another allotrope of Carbon which is an insulator with a large band-gap of 5.6eV as shown in Figure 9.

Diamond utilizes sp^3 hybrid orbitals to form σ-bonds in a tetrahedral structure as shown in Figure 9a. Valence band is completely filled up and conduction band is completely empty and bandgap is 5.6eV. Hence diamond is a perfect insulator as shown in Figure 9b.

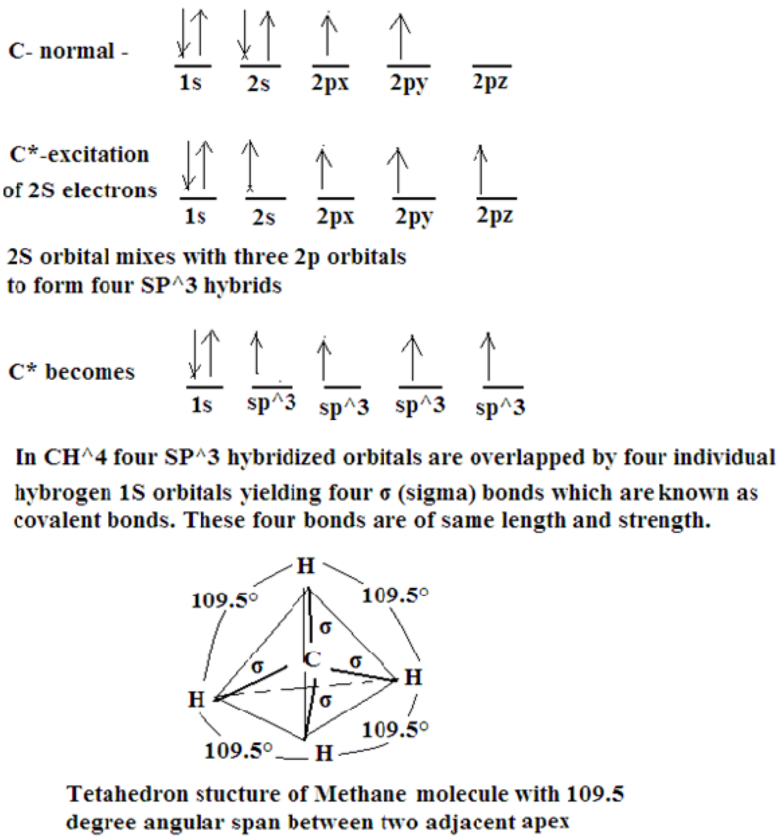
In Graphene, only two of the three available 2p orbitals form a total of three sp^2 hybrids with one 2p orbital remaining unhybridized. Two C atoms utilize one sp^2 hybrid each to mutually link up in a σ-bond or valence bond in a singlet coupled manner and ‘lone pair’ and $2p_z$ of the given two C atoms link up to form π-bond and ‘lone pair’. It is this $2p_z$ electron which is quantum tunneling from one sub-lattice to another and thereby contributing to the conductivity of the sample. Thus C-C have double bonds resulting in zero bandgap . The remaining two hybrids of each C atom overlap with the sp^2 hybrids of the neighbouring C atoms also in a singlet coupled manner and ‘lone pair’ as shown in Figure 8. Thus a benzene ring or hexagonal ring of C is formed in 2D layer extending to infinity in 2Dimensions.

So Carbon atoms in Graphene are arranged to form a honeycomb lattice tightly held by σ bonds between sp^2 hybrid orbitals. This gives rise to occupied σ band at energies well below Fermi-

Energy Level E_F and empty anti-bond σ* band at energies well above E_F as shown in Figure 8.

The remaining valence electrons(one 2p orbital for each atom) populate π-band which localize above and below the 2-D lattice with a node on the surface plane (see Figure 11).

This results in a robust honeycomb structure with E_F placed at 0 energy level.



2S orbital mixes with three 2p orbitals
to form four SP^3 hybrids

In CH^4 four SP^3 hybridized orbitals are overlapped by four individual hybrogen 1S orbitals yielding four σ (sigma) bonds which are known as covalent bonds. These four bonds are of same length and strength.

Tetahedron stucture of Methane molecule with 109.5 degree angular span between two adjacent apex

Figure 6a.Methane molecule formation through

four sp 3 hybridized orbitals of Carbon being overlapped by four individual Hydrogen 1s orbital.

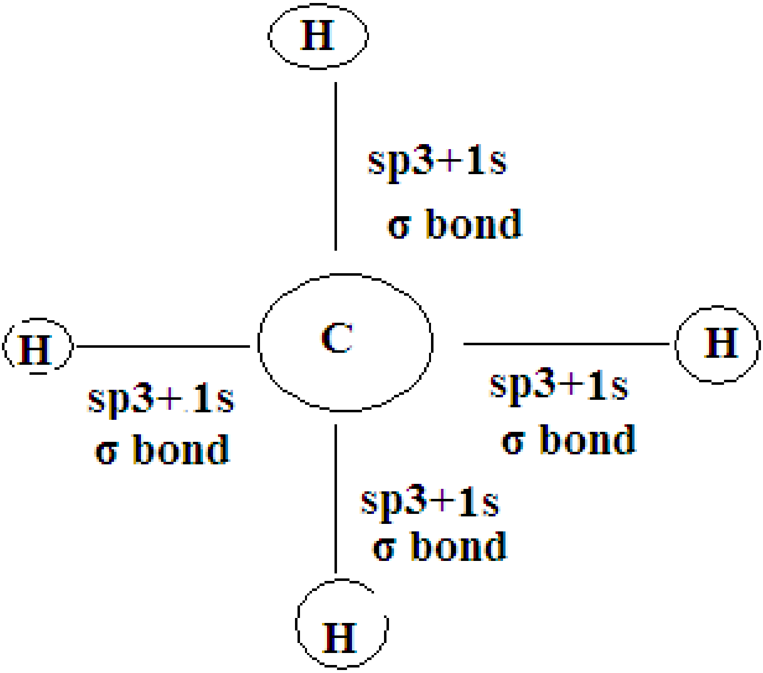


Figure 6b. The symbolic expression of Methane Molecule. Carbon atom is covalently bonded to 4 Hydrogen atoms through (sp 3 +1s) sigma bond.

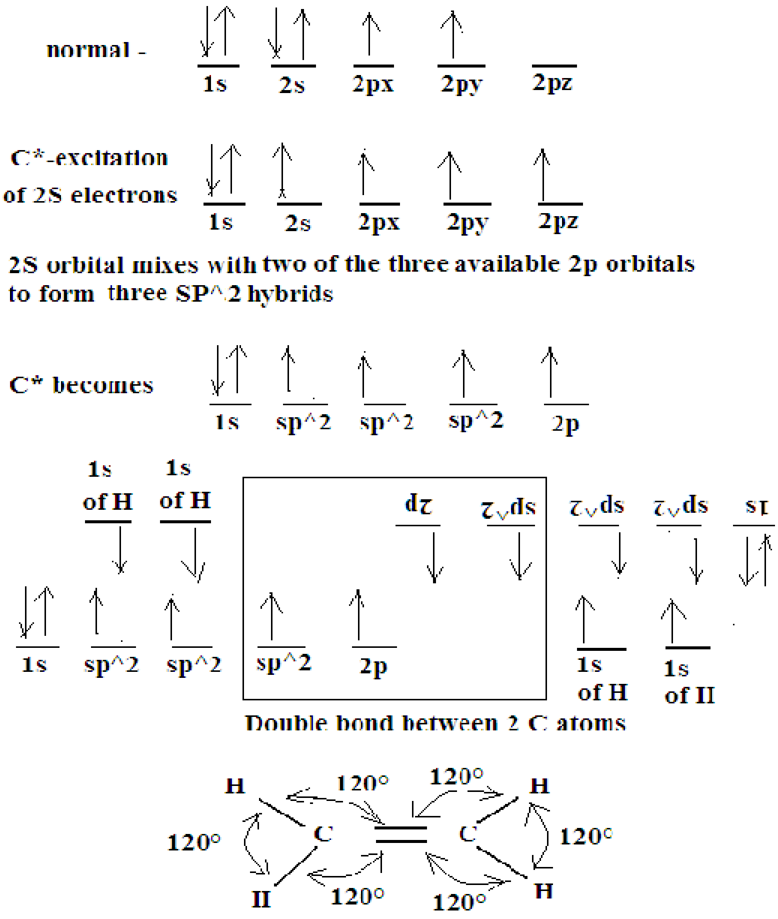


Figure 7a.Ethene molecule C 2 H 4 formation. There is a double bond between two C atoms and two sp 2 hybrid orbitals of each C atom's octave is completed by 1s orbital of two H atoms.

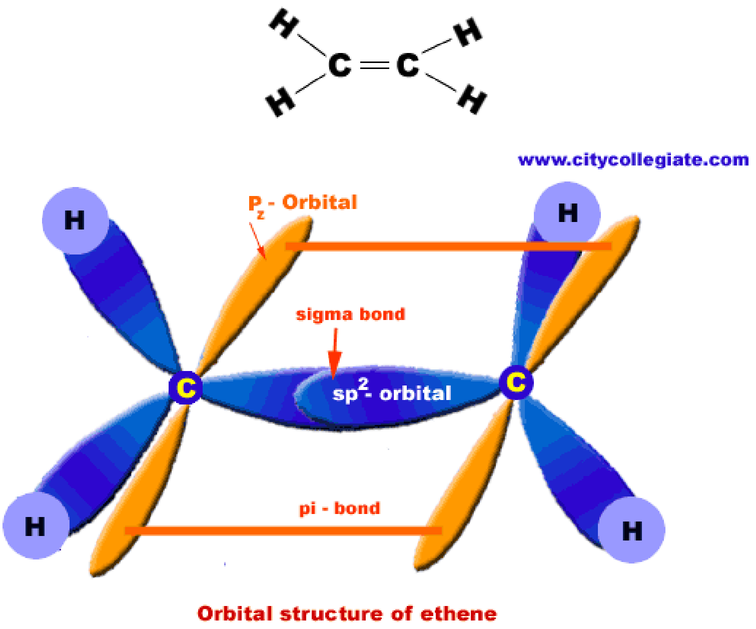


Figure 7b. The symbolic expression of Ethene molecule. The two sp 2 hybrid orbital of each C atom covalently bond with two H atoms and two C atoms form a sigma bond through sp 2 hybrid orbital and pi-bond through lone 2p orbital. Thus in Ethene two C atoms are double bonded and has zero band-gap due to σ and π bonds.

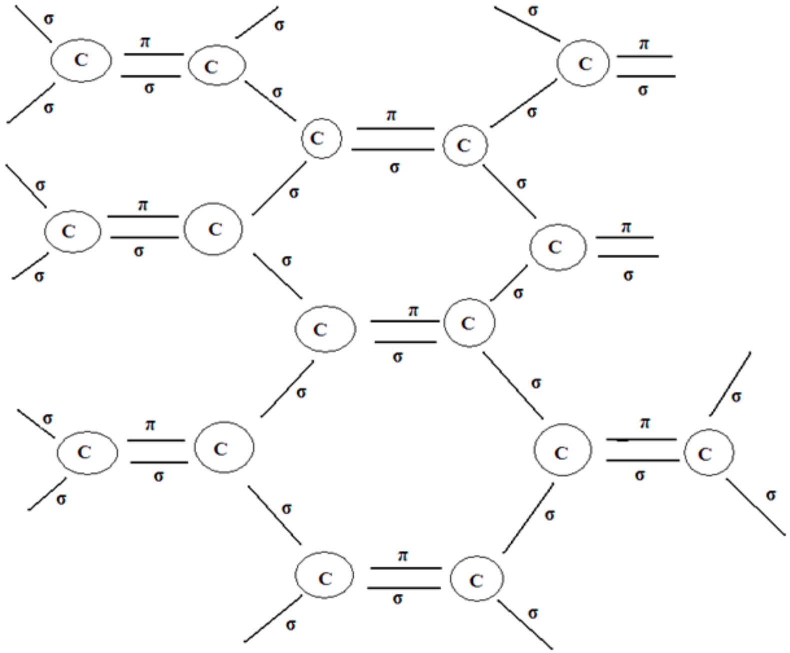


Figure 8. Graphene:One-atom-thick planar sheet of sp 2 - bonded and 2p-bonded carbon atoms through σ-bond and π-bond respectively. These are closely-packed in a honeycomb crystal lattice. C-C bond length is 1.42A° in a single graphene layer and in a stack of graphene which is graphite the interplanar spacing is 3.35A°.

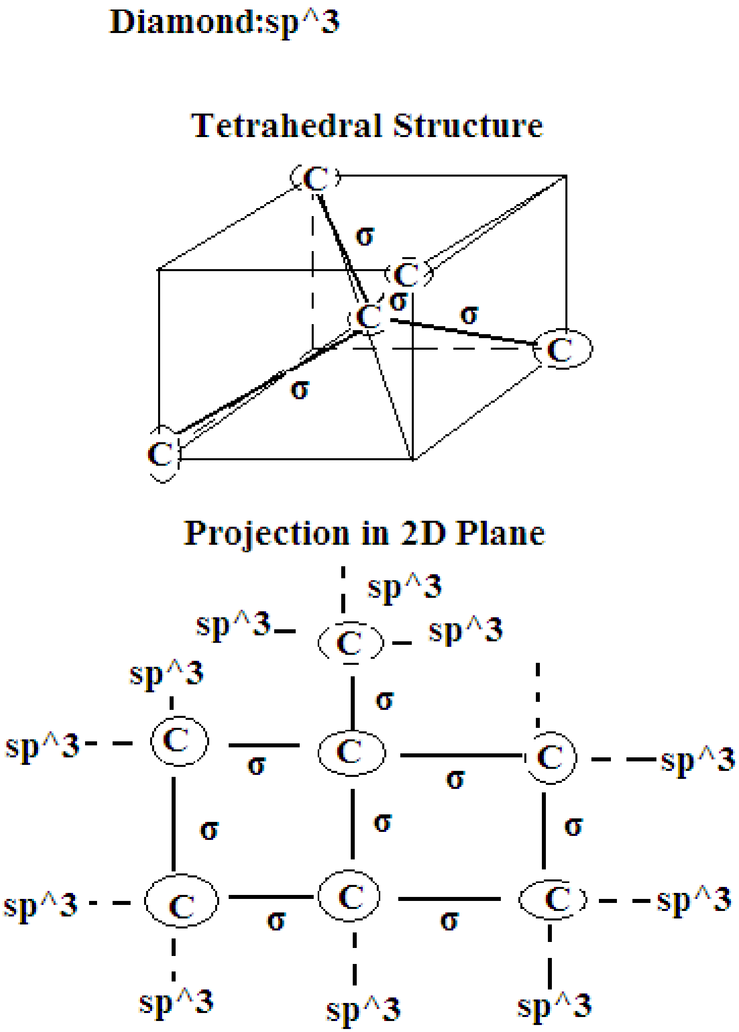
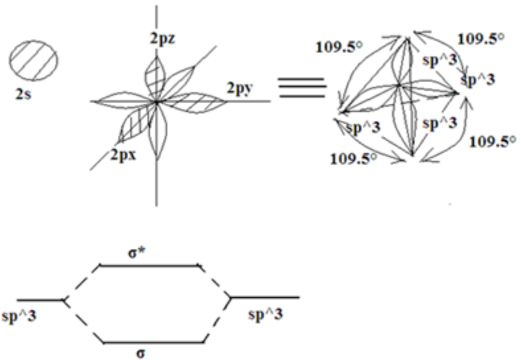
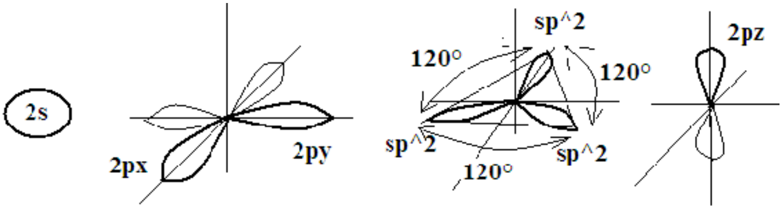


Figure 9a.Tetrahedral structure of Diamond and its projection in 2D plane.

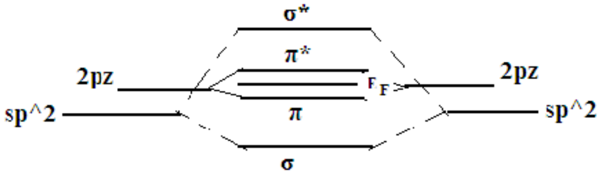


sp^3 octave completed in symmetric manner gives rise to σ bond which is filled up.
sp^3 octave completed in anti-symmetric manner gives rise to σ* anti-bond which is empty.
The energy difference between σ band and σ* band gives rise to band-gap of 5.6eV.
Hence Diamond is insulator.

Figure 9b. sp 3 hybridization results in a 3D tetrahedral geometrical structure which is the basis of Diamond tetrahedaral crystal with a band gap of 5.6eV.



Triagonal hybridization results in 2D sp^2 configuration directed towards the apex of an eqilateral triangle. The angular span between any two apex is 120°.
2pz orbit lies perpendicular to xy plane.



sp^2 symmetric and anti-symmetric overlap gives rise to bond σ band and anti-bond σ* band resulting in a large energy split between σ and σ* band. 2pz symmetric and anti-symmetric overlap result in a much closer energy split between π band and anti-bond π* band. σ band split and π band split result in a zero-bandgap semi-conductor as shown in the figure.

Figure 10. sp 2 hybridization results in 2D layer of C as shown in Figure 8 with zero band-gap.

There is an ‘anti-bonding’ π* band which is empty at T = 0K and charge neutral. The π-band and π*-band are much closer to E_F as shown in Figure 10. Hence this empty π*-band can easily be filled up with carriers by applying a gate potential in a typical FET configuration. Thus the state of a given 2-D graphene system is a superposition of these different ways of binding and this allows the system to stay in a distinctive energy minima. Such π/π* band system governs the low-energy (upto 2eV) behaviour of charge carriers in graphene and is responsible for most of the extra-ordinary properties of this material. Because of close proximity of π/π^* band system, we have zero bandgap metal/semiconductor.

This also results in a very low density of crystal defects in graphene prepared by mechanical exfoliation.

I.C.Chips of Future Generation Part 3.carriers-phonon interaction in Graphene.
Future IC Chips Part 3 supplement describes the special features of carrier-phonon interaction in Graphene which is an ideal testbed to demonstrate textbook condensed matter phenomena.

8.3. Supplement_Carrier-Phonon interaction in Graphene-an ideal test bed for demonstrating textbook condensed matter phenomena.

In conventional metals with large Fermi Sphere and Fermi Surfaces, Debye Temperature (θ_D) , the characteristic phonon energy scale, sets the demarcation between high temperature and low temperature behavior in electron-phonon scattering. This Debye temperature depends on the tensile strength of the material. Since graphene has the highest tensile strength hence its Debye Temperature is 2800 Kelvin. What this means is that above 2800K, all the phonon modes are excited namely Longitudinal Optical (LO) phonon, Transverse Optical (TO) phonon, Longitudinal Acoustic(LA) phonon and Transverse Acoustic(TA) phonon. That is in graphene, above 2800K, all phonon modes are excited and phonon population is proportional to absolute temperature and Resistivity (ρ) varies as T (Kelvin).

In large Fermi Sphere, below θ_D phonons start freezing out and $\rho(T) \sim T_5$ in 3-D metal and $\rho(T) \sim$

T^4 in 2-D metal. But this is not the case in graphene.

In case of metals with small Fermi-Sphere such as graphene and doped semi-conductors, all the phonon modes are excitable well below θ_D upto Bloch-Gruneisan(BG) temperature θ_{BG} where θ_{BG} is given as below [Efetov & Kim(2010)]:

$$\theta_{BG} = \frac{2\hbar v_S k_F}{k_B} < \theta_D \text{ where } \theta_D = \frac{\hbar v_S q_D}{k_B} \qquad (i)$$

$$Therefore\ \frac{\theta_D}{\theta_{BG}} = \frac{q_D}{2k_F} \qquad (ii)$$

For Graphene , which is unusually stiff, the Debye Temperature is 2800K and it has small Fermi Sphere hence $2k_F < q_D$. Here k_F is the radius of Fermi-sphere of electron in graphene or the Fermi wave vector given by Eq.(iii). Corresspondingly $\theta_{BH} < \theta_D$.

In Eq.(i), v_S = velocity of sound = 10^4m/s in graphene; q_D = maximum phonon wave vector > $2k_F$. Therefore θ_{BG} is significantly less than θ_D = 2800K.

Eq.(i) can be further simplified if we remember that carriers in graphene follow linear dispersion law as given in Eq.1 in Section 4.1. namely:

$$E_F(fermi\ Energy) = \hbar v_F k_F \qquad (iii)$$

In Eq.(iii), ħ = Planck's Constant/(2π); v_F = c/300 = 1.1×10^{6} m/s = Fermi Velocity;

Fermi Wave Vector = k_F;

Substituting Eq.(iii) in Eq.(i) we get:

$$\theta_{BG} = \frac{2\hbar v_S}{k_B} \times \frac{E_F}{\hbar v_F} = 2\frac{v_S}{v_F} T_F \qquad where\ T_F(Fermi\ Temperature) = \frac{E_F}{k_B} \qquad (iv)$$

In Eq.(iv) k_B = Boltzmann Constant, E_F = 0.11eV. Therefore θ_{BG} for graphene according to Eq.(iv) is 26Kelvin. But experimentally this comes to be 200K[Hwang & Sarma (2008)]. This is due to the uncertainty in the value of deformation-potential coupling constant (D).

Eq.(iv). also shows that Bloch-Gruneisen Low Temperature activity range (θ_{BG} = 26K theoretical) can be tuned by the gate voltage. Efectov et.al. (2010) have shown that graphene exhibits a very distinct crossover from linear T dependence to superlinear T^4 dependence well below Debye Temperature of 2800K and enters Low Temperature Activity range defined by Bloch and Gruneisen. By adjusting the gate voltage over ±1V, Bloch-Gruneisen Temperature was altered by 1 order of magnitude. Efectov and Kim (2010) also establish the universal scaling behavior of the normalized resistivity with normalized temperature. This represents 2D nature of electrons and phonons along with the chiral nature of carriers in graphene. Table

(i) shows one order of magnitude alteration in Bloch-Gruneisen Temperature Scale by tuning the Fermi Level of the sample.

Table i. Alteration of B-G Temperature scale by suitably tuning the Fermi-Level of the Sample.

Fermi Level of the sample (eV)	θ_{BG} (K)
0.1	21.1
0.2	42.2
0.3	63.5
0.4	84.4
0.5	105.5
1	210.1

As seen in Eq.(iv) by adjusting the gate voltage in gated Graphene FET (see Section 8.3.4.2.1 , Figure 12) and hence E_F by ±1V, Bloch-Gruneisen temperature (θ_{BG}) can be altered by one order of magnitude.

Hence Graphene is amazingly tunable:

1. E_F can be tuned by chemical doping [Schedin et.al.(2007), Chen et.al.(2008)];
2. E_F can be tuned by the gate voltage (Section

8.3.4.2.1);

3. K (dielectric constant- strength of electron to electron interaction) depends on the substrate on which graphene is epitaxially grown. This 'K' introduces an asymmetry to 'A' and 'B' sub-lattices which in turn introduces a band gap in the otherwise metallic graphene (Section 8.3.4.7).

At 300K, mean free path of electron in graphene is several microns, an order of magnitude better than that in semi-conductors because of reduced phonon-electron interaction at 300K and hence this leads to intrinsic resistivity $\rho(300K) = 30$ ohm [Chen et.al. (2007)]. This corresponds to $1 \times 10^{-6}\Omega$-cm in 3D metal which is lower than that of Silver , the most electrically conductive material till date. Graphene has the highest transparency [Nair et.al.(2008)]. Hence graphene is the ultimate transparent conducting electrode material. Graphene optical property as discussed in the foot note is another evidence for graphene being a relativistic system belonging to the realms of Quantum ElectroDynamic System.

[Foot-note:

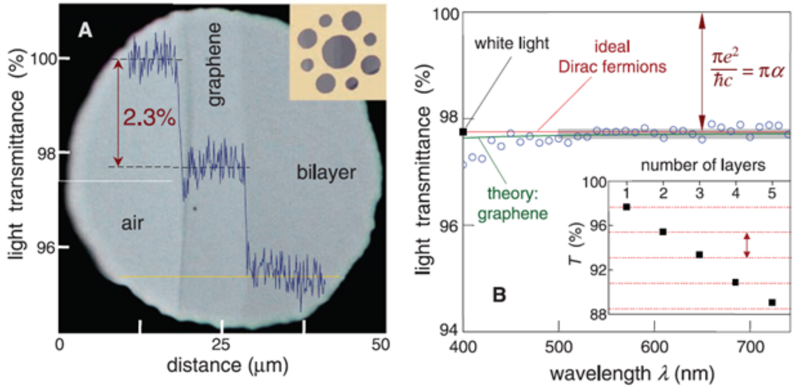


Figure A. Looking through one-atom-thick crystals. (A) Photograph of a 50-mm aperture partially covered by graphene and its bilayer. The line scan profile shows the intensity of transmitted white light along the yellow line. (Inset) Our sample design: A 20-mm-thick metal support structure has several apertures of 20, 30, and 50 mm in diameter with graphene crystallites placed over them. (B) Transmittance spectrum of single-layer graphene (open circles). Slightly lower transmittance for l < 500 nm is probably due to hydrocarbon contamination (5). The red line is the transmittance T = (1 + 0.5pa)–2 expected for two-dimensional Dirac fermions, whereas the green curve takes into account a nonlinearity and triangular warping of graphene's electronic spectrum. The gray area indicates the standard error for our measurements (5). (Inset) Transmittance of white light as a function of the number of graphene layers (squares). The dashed lines

correspond to an intensity reduction by pa with each added layer.

[adapted from Nair et.al.(2008)]

There are very few phenomena in condensed matter physics which are governed by fundamental constants only and have nothing to do with material parametrers. (h/e 2) is the resisitivity quantum and appears in Quantum Hall Effect and in Universal Conductance fluctuations. (h/(2e)) is the magnetic flux quantum and appears in super conductivity. Similarly opacity of a single graphene is:

$$\pi\alpha = \frac{\pi e^2}{\hbar c} = 2.3\% \tag{A}$$

As seen in Figure A(A), white light is absorbed in quantas of 2.3%. As number of intervening graphene layers increase so does the amount of absorbed light. It increases in step of 2.3% as shown in the inset of Figure A(B). The spectrum analysis shows that all the wavelength in the white light is absorbed in equal amount that is 2.3% for all practical purposes within observational errors.

Fine structure constant α = e 2 /(4πε 0 ħc) ~ (1/137) shows the coupling between light and relativistic electron and it is traditionally associated with Quantum ElectroDynamics and not with material science. That graphene opacity is governed by the fine structure constant proves that grapheme is a relativistic *system.]*

It is also found that electron-phonon scattering is independent of E_F and n_S (carrier concentration). At $n_S = 1\times10^{12}/\text{cm}^2$, mean free path is greater than 2 micron and this results in intrinsic mobility of 200,000 cm^2/(V-s). This is larger than InSb mobility at room temperature which is 77,000 cm^2/(V-s) and larger than that of Carbon Nano Tube (CNT) which is 100,000 cm^2/(V-s).

The kind of massless electron we see in super-conductors at liquid He temperature we witness the same kind of kinematics at 300K in Graphene. Hence graphene provides a rare opportunity for doing quantum mechanical research.

Graphene is highly simplified and conceptually easy to visualize. Dirac Equation is obeyed by massless relativistic carriers even at Room Temperature. Psuedospin degree of freedom plays the role of relativistic spin(Section 8.3.4.2.6). Dispersion of Electronics Band is described by a single parameter v_F (Fermi Velocity) as seen in Eq.1 and also in Eq.iii.

The combination of simplicity and tunability makes graphene an ideal test bed for demonstrating fundamental condensed matter phenomena.

It has 100× tensile strength as compared to that of Steel. It has much better thermal conductivity than that of Diamond which till now was the most thermally conductive material. It is also most closely packed structure. It is impermeable to the smallest atom like Helium but it still can stretch.

It is zero band-gap material meaning by conduction band and valence band lie close together hence graphene is a conductor but it cannot switch on and off.

In semi-conductors, because of the band- gap, photon must exceed the bandgap to cause photo-excitation and absorption hence semi-conductors are susceptible to only certain part of the spectrum. In graphene there is no such limitation. Graphene has equal absorption coefficient from UV to far IR light because it has zero bandgap.

Exploiting Graphene's optical and electron abilities along with strength and flexibility could lead to foldable plastic smartphones, cheaper solar cells or sensors that can detect single molecule of gas or identify individual DNA bases.

Silicon took decades to find its signature role in technology. Graphene's journey has just begun.

Graphene has superlative mechanical, thermal and electronic properties as shown in Table 3[Savage (2012)].

The Integrated Circuit Technology of Future Generations-Part 3_mobility measurement, defect estimate, balllistic nature, conductivity variation. Future IC Chips_Part 3, describes Quantum Electrodynamic properties of Graphene. It gives the mobility measurement and defect estimate results. It gives the possibility of achieving Ballistic Transistor at room temperature and it gives the gate control of conductivity.

8.3.3 Electrical Mobility measurement and Defect Estimation.

A mobility of $200{,}000 cm^2/(V\text{-}s)$ is achieved in suspended graphene samples with electron density of $2 \times 10^{11}/cm^2$. From this mobility it is estimated that the upper bound on the native crystalline defect density of exfoliated graphene is $6 \times 10^9/cm^2$ [Bolotin et.al. (2008)].

This residual crystal defects are a potential source of intervalley scattering. Crystalline defects are present in various concentrations in graphene synthesized by reduction of graphene oxide, by CVD or by segregation of C on SiC.

The pristine graphene has metallic behavior. But even a small amount of irradiation drastically affects the low temperature behaviour.

In pristine graphene, minimum conductivity is temperature independent from 4 to 100K. But irradiated graphene is insulating with diverging resistivity at T approaching 0 K.

Electrons in Graphene behave like particles with zero rest mass, known as chiral Fermions or neutrinos [Novelselov et.al. (2005)]. Electrons are best described by Relativistic Quantum Physics. Graphene is condensed matter system. [In condensed matter system we study the physical properties of condensed phase of matter. Solid and Liquid are ordinary condensed phases. Super conductors and Bose-Einstein Condensate are exotic condensed phases.] In graphene, electrons mimic zero-mass particles governed by Dirac's (Relativistic) equations and have effective speed of light ~ 10^6 m/s = c/300. Hence the effective mass of conducting electron in Graphene is:

$$m(effective\ mass) = \frac{m_0(rest\ mass)}{\sqrt{1-(\frac{v_0}{c})^2}} \qquad 2$$

Where m_0 = zero rest mass and v_0 = 10^6 m/s = Fermi velocity of electron mobile carriers.The effective mass is experimentally measured as $0.007m_e$ [Zhang et.al.(2005c)].

The 'holes' in graphene are the quantum electrodynamical equivalent of prositrons. Hence electrons and holes are charge conjugate symmetrical particle described by the same Dirac

Spinor Function.

8.3.4. Quantum-Electrodynamical behavior of electron carriers confined in 2D graphene layer [Geim & Novoselov (2007)] .

8.3.4.1. The possibility of achieving Ballistic Transistors(the holy grail of Electronics Engineers) through the fabrication of GrapheneFET(GFET)

Charge carriers in graphene have an unique nature. In condensed matter physics the Schrodinger Equation is sufficient to describe the electronic properties of materials. But in graphene the charge carriers mimic relativistic particles. Hence carriers in graphene behave like quasi-particles and are described by the Dirac Equation.

Electrons interaction with periodic potential variation of graphene honeycomb lattice gives rise to a new quasiparticle that at low energies are described by (2 Space + 1 Time)-dimensional Dirac Equation with an effective speed of $v_0 = 10^6$m/s. These massless Dirac Fermions can be seen either as electrons which have lost their rest mass or as neutrinos which have acquired a charge ‘e’.

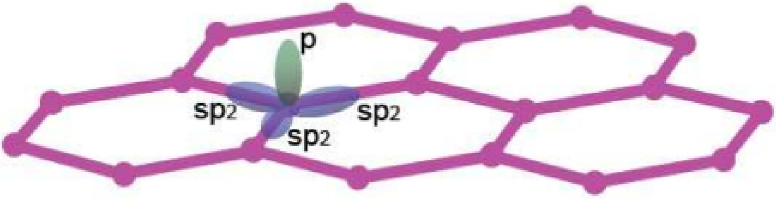


Figure 11.The graphene honey-comb lattice with C atom placed at the lattice sites. Hybrid sp 2 symmetric overlap with the neighbouring hybrid sp 2 gives rise to the covalent “backbone” of graphene with honeycomb lattice geometry. The remaining 2p z orbital symmetric overlap with the available 2p z orbital creates half filled π-bond. It is this half-filled π-bond which provides free electrons that tunnel from site to site giving rise the fascinating electronic properties of graphene [Pachos (2009)]

Graphene is a large molecule of C atoms where C atoms are covalently bonded in a flat configuration. One C atom occupies one lattice center. Three(sp2 electrons) of the four valent electrons of each C atom are covalently shared with the corresponding valent electrons(sp2) of the three neighbouring C atoms via σ bonds which gives rise to completely filled σ energy band. These electrons are highly localized and tightly bound to the host C atom and donot contribute to mobile carriers or conductivity of the graphene. The σ-bond provide the covalent “backbone” structure of graphene with honeycomb lattice geometry. Their strength is responsible for the flexibility and robustness of the 2D graphene lattice though 3D graphite is very soft.

The $2p_z$ orbital of the fourth electron symmetrically overlap the available unpaired $2p_z$ orbital of one of the three neighbouring C atoms to create a π-bond which give rise to half-filled π energy band. Each π-bond allows its $2p_z$ orbital electron to tunnel from one atom to the neighbouring atom. Hence graphene is a many body system: several electrons are allowed to tunnel from site to site while at the same time satisfy Pauli-Exclusion Principle.

Each carbon atom effectively contributes 1 electron in $2p_z$ orbital for site to site tunneling which creates the electrical conductivity and there are two atoms per unit cell of hexagonal shape. The area of one unit cell is:

$$\begin{aligned} &Total\ area\ of\ one\ Unit\ Cell = A \\ &\quad = [d \times 2dCos30] + 2\left[dSin30 \times \frac{2dCos30}{2}\right] \qquad 3 \end{aligned}$$

Where d = lattice constant = 1.42A°

Therefore A(total area) = 5.238760873 (A°)^2 = 5.238760873 $\times 10^{-20}$ m^2

So the aerial density per cm^2 = $(1 \times 10^{-4})/(5.238760873 \times 10^{-20})$ = $1.908848341 \times 10^{15}$ unit cells per cm^2

Each unit cell has two carbon atoms hence maximum carrier concentration possible

$$n_{max} = p_{max} = 3.8177 \times 10^{15}\ carriers\ per\ cm^2$$

Simple geometrical configuration of C atoms gives graphene a surprisingly rich collective behavior at microscopic level. They exhibit non-zero Berry Phase which prevents back-scattering [Section 4.1]. This enables the mean free path of electron in micron or sub-micron range at room temperature. The mean free path is independent of carrier concentration and doping density and weakly dependent on ambient temperature hence the mobility is unusually high . Routinely we get a mobility greater than 15000cm2/(V-sec) and with extra care mobility as high as 200,000cm2/(V-sec) can be achieved. From this mobility the mean free path can be calculated from Eq.4 below [Bolotin et.al.(2008)] . In this equation 'l' is the mean free path of the carrier, 'n' is the carrier concentration, 'e' is the electron charge and 'h' is Planck's constant.

$$\sigma = en\mu = \frac{2e^2}{h}\left[(\pi n)^{\frac{1}{2}} l\right] \tag{4}$$

Rearranging Eq.4 we get the mean free path as:

$$l = \frac{\mu h}{2e}\sqrt{\frac{n}{\pi}} \tag{5}$$

Table 4 tabulates the value of mean-free path of carriers for different carrier concentration and different motilities'.

Table 4. Mean free path of the carriers in graphene for different mobility and different carrier concentration at 300K.

μ	15,000cm^2/(V-sec)	70,000cm^2/(V-sec)	200,000cm^2/(V-sec)
n	l(μm)	l(μm)	l(μm)
10^{11}/cm^2	0.05	0.258	0.74
10^{12}/cm^2	0.175	0.8	2.33
10^{13}/cm^2	0.55	2.58	7.4
10^{14}/cm^2	1.75	8.166	23.3
10^{15}/cm^2	5.534	25.8	73.8

From the Table 4 it is evident that carriers are in ballistic mode and we should be able to fabricate ballistic transistors at room temperature which has been the ultimate wish list of Electronic Design Engineers.

Ballistic Transistor is defined as a device where channel length is much shorter than the mean free path of the carrier which is the case in graphene transistor.

Hence transit angular frequency of an ideal graphene transistor is as follows:

$$\omega_T = 2\pi f_T = \frac{1}{\tau_{transit}} = \frac{v_F}{L} \qquad 6a$$

Therefore transit frequency of an ideal ballistic transistor is:

$$f_T = \frac{v_F}{2\pi L} \qquad 6b$$

But we donot have an ideal graphene transistor. In real life graphene transistor we have metal contacts at Source and Drain and in metal contact we have non-relativistic Schrodinger electrons and holes. Thus the Graphene FET has mixed-particle entity: within source Schrodinger particles obeying electric drift kinematics are taking part in transistor action; along the channel length the non-relativistic Schrodinger particle is transformed into relativistic Dirac quasi-particles and again at the Drain relativistic Dirac quasiparticles transform into non-relativistic Schrodinger particles. This lowers the Fermi velocity at which ballistic action takes place. Also the Transmission at normal incidence in Source is less than Unity. It is because of these two reasons that the cut-off frequency experimentally observed in gated ballistic transistors is much lower than the prediction of Equation 6.

D.Dragoman and M . Dragoman (2011) have carried out numerical calculation which confirms that ballistic action is carried out at almost half the Fermi-velocity. So the advantage of Dirac quasi-particles is not fully derived in metal contact GFETs as shown in Table 5.

Table 5. Improvement in transit frequency of ideal and real ballistic transistors as the channel length is scaled down as given by Eq.6 in the text.

Channel length(nm)	Ideal GFET f_T	Real GFET f_T
100	1.6THz	795GHz
75	2THz	1THz
50	3THz	1.6THz
25	6.4THz	3.2THz
10	16THz	8THz
5	32THz	16THz
1	159THz	79.6THz

Table 5 clearly brings out the possibility that gated 2-D graphene systems could potentially become multifunctional high speed and high power transistors, thus introducing a prospective paradigm shift in future generation of micro-(and nano-)electronics. [Hwang *et.al.(2007)]*

8.3.4.2. Quantum-mechanically enhanced carrier transport phenomena in Graphene.

Dirac Electrons in 2-D graphene are markedly different from the Schrodinger Electrons in Quantum Well as seen in Semi-conductor Photonic Devices.

Each Carbon atom in graphene honeycomb lattice forms strong covalent bond with neighboring three Carbon atoms through the symmetric overlap of its three sp_2hybrid orbital. One $2p_Z$ is left over to wander across the two dimensional crystal.

Electron motion in Graphene is equivalent to that of a neutrino or a relativistic Dirac Electron with vanishing mass. This causes a nontrivial Berry Phase under 2π rotation in wave vector space leading to the absence of backscattering and in CNT perfect conduction even in the presence of scatterers.

In CNT, circumferential chiral vector determines the fictitious Aharonov-Bohm flux.CNT becomes metallic when flux vanishes and semi-conductor when flux is non-zero. Another way of looking at this is ‘quantum coherence’ . If quantum coherence of electron motion around the circumference of CNT is maintained then it is metallic otherwise it is semi-conducting [Zheng et.al. (2002)]. In contrast conductivity of graphene is independent of E_F (

Fermi Energy) and n (electron concentration) as long as variation in effective scattering is neglected. Therefore Graphene should be treated as a metal and not zerogap semiconductor.[Ando (2009)]

In contrast 3-D graphite is a semi-metal whose bands slightly overlap and allow pockets of electrons and holes to tunnel between staggered layers. Here the carriers have a quadratic energy dispersion relationship as given in the equation below:

$$E = \frac{|\hbar k|^2}{m^*} \tag{7}$$

m⋆ is the effective mass which accounts for the carrier interaction with the periodic potential distribution throughout the crystalline lattice due to the orderly arrangement of the lattice sites.

In a single 2-D graphene sheet, the overlap shrinks down to a single point K and K′ also known as Dirac Points. The two Dirac Points correspond to the two symmetric sub-lattices A and B of the 2-D graphene crystal (see Figure 17). This results in two cones aligned longitudinally and just touching at the apex as shown in Figure 5b. At low energies, this results into linear energy dispersion relationship as given by Equation 1.

Because of the linear energy dispersion relationship, electron dynamics is not modeled by Schrodinger Particle but instead by Dirac's Quasi-particle. The dispersion relationship implies that electron's mass vanishes through a large range of momentum values in the crystalline lattice.

Electrons are not massless but their transport properties are more like those of PHOTON.

In gated graphene as shown in Figure 12a, increased gate voltage raises the Fermi Level E_F into the conduction band and negative voltage lowers the Fermi Level into the valence band as shown in Figure 15. Positive Voltage results into the enhancement of 'n'(electron concentration) and negative voltage results into enhancement of 'p'(hole concentration).

8.3.4.2.1. Conductivity Variation in Gated Devices.

Just as in enhancement mode and depletion mode MOSFET the channel carrier density can be controlled by the gate voltage in exactly the same manner with a proper gated structure as shown in Figure 12a, the concentration of carriers in graphene layer can be precisely controlled by the gate voltage. The concentration variation profile is given in Figure 12b. It should be noted that the maximum concentration of electron or hole achieved in Graphene ($\sim 5 \times 10^{13} cm^{-2}$) is comparable to the channel carrier concentration induced in a comparable MOSFET.

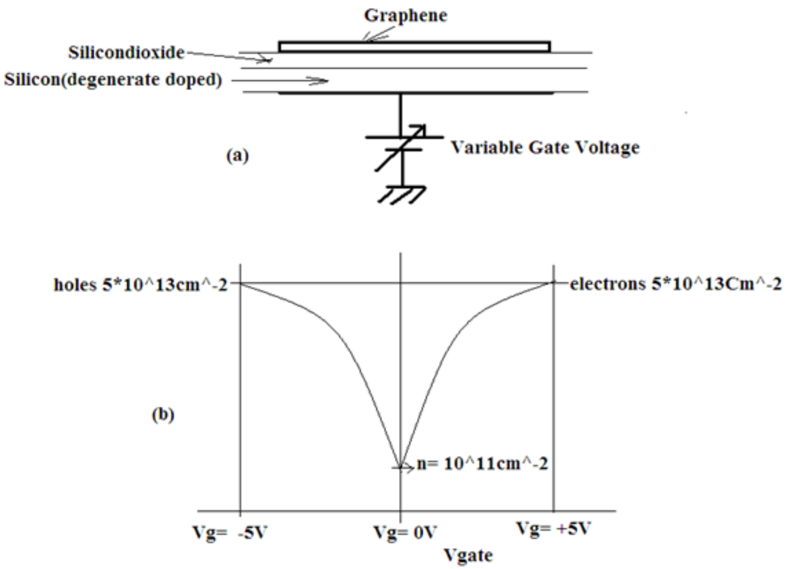


Figure 12. In upper figure (a) the gated structure of graphene is given. In lower figure (b) the hole and electron concentration variation with respect to gate voltage is shown.

From Figure 12 it is evident that graphene exhibits pronounced ambipolar electric field effect namely electron and hole can be tuned continuously with gate voltage from $10^{11}/cm^2$ to $10^{13}/cm^2$. Mobility is weakly dependent on temperature T. This means mobility in Graphene is limited by impurity scattering and hence can be significantly improved.

In InSb only under intrinsic condition mobility is as high as 77000 $cm^2/(V\text{-}s)$. But in graphene mobility remains high even at highest electric field-induced concentration and seems to be little effected by chemical doping. This translates into ballistic devices at submicron level (shorter than 0.3micron channel length at 300K) as already discussed in Section 4.4.1. Room temperature Ballistic transistor has been the holy grail for Electronics Engineers and graphene should make it happen. Large value of Fermi Velocity and low resistance contacts without schottky barrier should reduce switching time to 10^{-13}s.

Quantum Theory applied to 2D graphene in the presence and absence of a magnetic field for two different cases of scattering, i.e. short and long range scattering, give results which are qualitatively different from those obtained by the application Boltzmann Transport Theory [Shon et.al.(1998)]. In high magnetic field limit, the conductivity exhibits a series of peaks whose values depend on natural constants and Landau Level Index. At E(energy) = 0 we have an universal conductivity $(e^2/(\pi^2\hbar))$ independent of magnetic field.

Long range scatterers are random charged impurity centers in the substrate at distance ‘d’ Angstrom from the interface. The short range scatterers are lattice defects in the graphene. Quantitative agreement between theory and experiment indicates that the dominant scattering mechanism is Coulomb scattering by random charged impurities located near the interface of graphene and substrate on the substrate side.

The typical impurity concentration in SiO_2 substrate is $n_i = 10^{12}/cm^2$. This corresponds to μ (300K) = 15,000 $cm^2/(V\text{-}s)$. If n_i can be reduced to $10^{10}/cm^2$ then μ would improve to 2×10^6 $cm^2/(V\text{-}s)$.

Theoretically at gate voltage = 0V, Fermi Energy is at 0eV and Density of State, $D(E_F)$ should be zero.But in real graphene devices because of charged impurity cwenters the density of state at Dirac Point is not zero. The presence of charged impurities causes potential fluctutations in otherwise symmetric and periodic potential distribution in the graphene 2-D lattice. Theoretical calculations including level broadening effect performed prior to experiment showed that conductivity takes an universal value:

$$\sigma_{min} = g_s g_v \frac{e^2}{2\pi^2\hbar} \quad at\ E_F = 0 \quad [Shon\ et.al.(1998), Novoselov, Geim\ \ et.al(2005)] \quad (A)$$

These charged impurities lead invariably to large scale density inhomogenities in 2-D carrier system. The carrier density fluctuation $\Delta n > n_s$ at low carrier density. Hence the system breaks into a random network of 2-D electron and holes conducting puddlesa, producing a finite conductivity at zero gate voltage.

Ideally at $V_G = 0$ V, $n_s = 0$. But graphene doesnot remain a homogenous zero gap semiconductor but a spatially inhomogenous semi-metal with small random puddles of electrons and holes 2-D liquid depending on the details of charged impurity configuration in the substrate.

Ideally Eq(1) holds for linear energy dispersion. At $V_G = 0$ V, $n = p = 0$.

At $T = 0$ Kelvin also $n = p = 0$. But in real life there are remote interfacial charge impurities . This causes electron and hole puddles at low carrier density.

Gate Voltage V_g causes an electron concentration:

$$n = \frac{K_{OX} V_g}{4\pi t} = \frac{K_{OX}}{4\pi t} \times \frac{E_F}{e} \qquad (B)$$

Rearranging the terms we get:

$$\frac{E_F}{e} = \frac{4\pi te}{K_{OX}} \quad \text{where } t = \text{oxide thickness and } K_{OX} = \text{dielectric constant of Oxide} \qquad (C)$$

At large gate voltage, there are large carrier concentration ‘n’ $>>$ Δn (due to potential fluctuations).

Hence at $n > n_i$, system is homogenized and

Boltzmann Transport holds good.

At $n < n_i$, the onset of the inhomogeneous puddles of electron and hole carrier starts and Boltzmann Transport Theory breaks down.

Experiment show that n_c (critical density of carriers) $= n_i \sim 10^{12}/cm^2$ is the characteristic density below which the onset of inhomogenity starts.

The fluctutation

$$\Delta n = \frac{n_i}{8\pi d^2} \quad where\ d = distance\ of\ impurity\ plane\ from\ graphene\ plane \qquad (D)$$

Hence transport near the Dirac Point is dominated by a spatially inhomogenous carrier density.

Hwang et.al. have shown that charge impurities in the substrate near the interface are the dominant source of scattering. In low carrier density regime that is $n < 10^{12}/cm^2$, carier density fluctuation caused by charged impurities leads to electron-hole puddles. This leads to σ_{min} even at gate voltage = zero. Finally the Fermi Temperature of 2-D graphene is 1300K for $n = 10^{12}/cm^2$ hence conductivity (σ) has no temperature dependence from 0 to 300Kelvin. It is limited by charged impurity scattering This is in excellent agreement with experiment.

The density of states in graphene is given as follows:

$$D(E) = \frac{|E|}{2\pi\hbar^2 v_F^2} \quad where\ v_F = Fermi\ velocity \tag{E}$$

Einstein relation gives the conductivity as:

$$\sigma_0 = g_v g_s e^2 D^* D(E_F) \tag{F}$$

In Eq.(E), $g_v = 2$ = valley degeneracy corresponding to the two symmetric sub-lattices.

D^* = diffusion coefficient, $g_s = 2$ = spin degeneracy.

Let τ = relaxation time due to impurity scattering. Then diffusion coefficient is :

$$D^* = v_F^2 \tau \tag{G}$$

Reciprocal of relaxation time is defined as follows(this relaxation time is different from mean free time):

$$\frac{1}{\tau} = \left(\frac{2\pi}{\hbar}\right) n_i < U_i^2 >_{E_F} \ D(E_F) \tag{H}$$

Where n_i = impurity density, U_i = matrix element of the impurity potential between initial and final states.

$<\ldots>_{Ef}$ = denotes the average potential at E_F.

Substituting Eq. (E), Eq.(G) and Eq.(H) in Eq.(F) we get:

$$\sigma_0 = g_v g_s e^2 v_F^2 \tau \frac{|E_F|}{2\pi\hbar^2 v_F^2} \qquad (I)$$

Simplifying Eq.(I) we get :

$$\sigma_0 = g_v g_s e^2 \tau \frac{|E_F|}{2\pi\hbar^2} \qquad (J)$$

Eq.(H) can be re-written as:

$$\tau = \frac{1}{\left(\frac{2\pi}{\hbar}\right) n_i < U_i^2 >_{E_F} \; D(E_F)} \qquad (K)$$

Substituting Eq.(K) in Eq.(J) we get:

$$\sigma_0 = g_v g_s e^2 \frac{1}{\left(\frac{2\pi}{\hbar}\right) n_i < U_i^2 >_{E_F} \; D(E_F)} \frac{|E_F|}{2\pi\hbar^2} \qquad (L)$$

Substituting Eq.(E) in Eq.(L) we get:

$$\sigma_0 = g_v g_s e^2 \frac{1}{\left(\frac{2\pi}{\hbar}\right) n_i < U_i^2 >_{E_F} \frac{|E_F|}{2\pi\hbar^2 v_F^2}} \frac{|E_F|}{2\pi\hbar^2} \qquad (M)$$

Define a dimensionless parameter ‘W’ which characterizes the strength of impurity scattering. We define it as follows:

$$W = \frac{n_i < U_i^2 >_{E_F}}{4\pi^2\hbar^2 v_F^2} \qquad (N)$$

Eq (M) is simplified as:

$$\sigma_0 = g_v g_s e^2 \frac{1}{\left(\frac{2\pi}{\hbar}\right) n_i < U_i^2 >_{E_F} \frac{1}{v_F^2}} \qquad (O)$$

Multiplying the numerator and denominator of Eq. (O) by ‘W’ we get:

$$\sigma_0 = g_v g_s e^2 \frac{1}{\left(\frac{2\pi}{\hbar}\right) n_i < U_i^2 >_{E_F} \frac{1}{v_F^2}} \times \frac{1}{W} \times \frac{n_i < U_i^2 >_{E_F}}{4\pi^2\hbar^2 v_F^2} \qquad (P)$$

Cancelling the common terms and simplifying Eq. (P)

$$\sigma_0 = g_v g_s e^2 \frac{1}{\left(\frac{2\pi}{1}\right)\frac{1}{1}} \times \frac{1}{W} \times \frac{1}{4\pi^2\hbar} = \frac{g_v g_s e^2}{4\pi^2\hbar W(2\pi)} \qquad (Q)$$

Ando et. al.(2009) give this the conductivity as :

$$\sigma_0 = \frac{g_v g_s e^2}{2\pi^2\hbar W} \qquad (R)$$

Eq(R) tells us that the conductivity is independent of E_F and carrier concentration ‘n’ as long as the strength of scatterer ‘W’ dependence on E_F and ‘n’ is neglected.

The carrier transport in Graphene has some unique properties. Graphene’s conductivity never falls below a minimum value corresponding to the quantum unit of conductance even when concentration of charge carriers tend to be zero. This quantum unit of conductance is [e2/(π2ħ)] [Ando(2006)]. This quantum unit is given by natural constants and is independent of scattering strength and band parameter. Theoretically it has been derived using Self-consistent Born Approximation of Quantum Theory [Shon and Ando (1998)].The measured minimum conductance is larger than the predicted value. According to Bolotin et.al(2008) as described by Eq.4 in Section 4.4.1 it is given as :

$$\sigma(min) = en\mu = \frac{2e^2}{h}\left[(\pi n)^{\frac{1}{2}} l\right] \tag{4}$$

$$\begin{aligned} where\ n = 2 \times 10^{11} cm^{-2}\ and\ l = 1.2\mu m \\ = mean\ free\ path \end{aligned}$$

It is further observed by Ando(2006) that electrical conductivity varies linearly with carrier concentration which means that mobility is independent of scatterers. No saturation is observed at high carrier concentration.

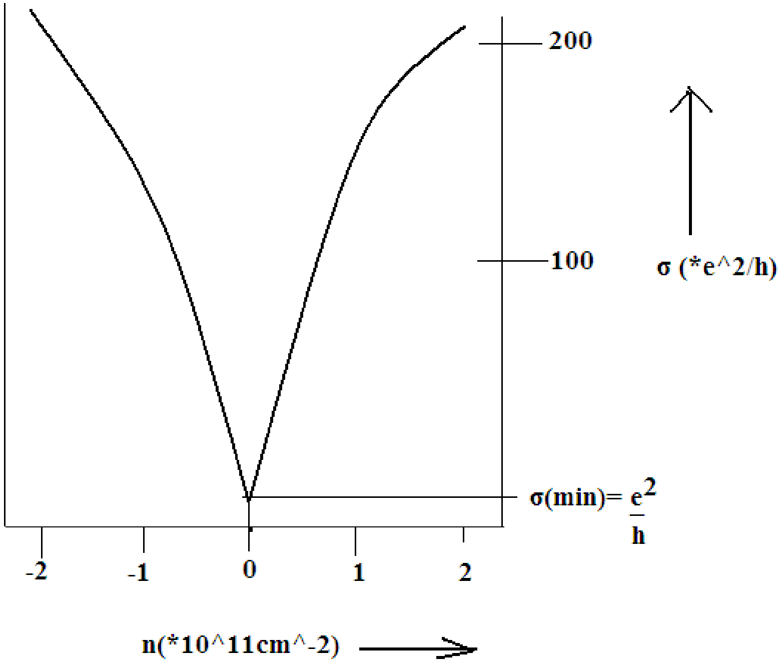


Figure 13. Sub-linear dependence of conductance on carrier concentration.σ

Bolotin(2008) has observed sublinear dependence of conductance in suspended graphene as shown in Figure 13. The sub-linear dependence shows that short range scatterers are involved . This may be associated with point defects or sample edges. The shape of conductivity indicates that high mobility sample contain short range disorder whereas low mobility sample are dominated by long range scatterers [Tan et.al.(2007)] . The scattering between K and K′ points of the energy band diagram in Figure 5a determine the range of scattering. If the range is shorter than lattice constant then we have short range scattering and if the range is comparable to or longer than the lattice constant but much smaller than typical electron wavelength then we have long range scattering. It can also be stated that Scattering between the two Dirac points implies long-range scattering and absence of scattering implies short range scattering [Zheng et.al. (2002)]. Point defects in graphene can be categorized as short and long-range scatterers. Flourination of the graphite surface causes a kind of local π-electron defect. This is a short range scatterer. When graphene is exposed to ozone and Ultra-Violet radiation then adsorption is caused. This adsorption is long range scattering potential [ibid]. A charged impurity or Coulomb impurity situated in the substrate near the graphene is long range scatterer. Some points have both short and long-range scattering characteristics. For example Boron atom in boron-doped graphene brings about a notable correction on the electron density only in the range of one unit cell around it hence this is a short-range scatterer. However boron atom slightly deforms the flatness of the graphite surface in a much larger range. Hence this is long range scattering of electron motion[ibid].

Graphene has all its bonds(three sp_2-σ bonds and one $2p_z$-π bond) completely saturated hence underlying substrate should have little or no effect on its conductivity. But Aizawa et.al.(1990) have experimentally found that transition-metal-carbide substrate has considerable bonding with the overlying graphene layer leading to considerable remote interfacial phonon scattering.

Bolotin(2008) have also proved that suspended graphene layer experiences more than an order of magnitude improvement in its mobility as compared to that of a sample overlaid on a substrate.

Both these experiments clearly prove that remote interfacial phonon scattering has a significant role to play in carrier transportation.

The Integrated Circuit Technology of Future Generations-Part 4_QED properties continued
IC of future generation_Part 4 is a continuation of Quantum Electro-Dynamic properties of Graphene.

8.3.4. Quantum Electrodynamical Properties of Graphene continued.

8.3.4.2.2. Vanishing Effective Mass of the carriers in graphene leads to Dirac quasi-particles

We have seen that electron in unbounded free space can assume all energy states but electron in an infinite potential well is permitted in only certain discrete states. The same electron in a solid is permitted bands of energy states and thus is born the subject of the band theory of Solids. The valence electrons are localized and bonded to their host nuclei but the conduction electron wave function spreads throughout the crystalline lattice and the crystalline lattice creates a periodic potential distribution which modifies the carrier wave-function. Electron moving through a 3D periodic potential distribution has a mass different from that of the free space electron. This mass of electron in a crystalline lattice is known as effective mass and is given by the formula[Sze (2012)]:

$$m_e^* = \frac{\hbar^2}{\frac{\partial^2 E}{\partial k^2}} \qquad 7$$

The magnitude and the sign of the second derivative of E-k diagram decides the mass and the nature of the particle moving through the lattice structure.

Figure 14 contains the profile of reduced E-k, dE/dk-k, d2E/dk2-k and m*-k diagrams. Inspection of m*-k diagram shows that electron has the lightest mass at the bottom of reduced energy band diagram and heaviest at around k(wave vector) = ±π/(2a) where a = lattice constant. Hence metals with Fermi level at k = ±π/(2a) are heavy fermion metals. Electrons at the top of the valence band has (-)ve mass which gives rise to the phenomenological particle ‘Hole’.

In graphene as we have seen in Eq.1. the energy dispersion relation is as follows:

$$E_{sk} = s\gamma k \qquad 1$$

$$\text{where } s = +1 \text{ for conduction band and } s = -1 \text{ for valence band}$$

$$\text{and } k = \text{wave vector} = |k| \text{ and } \gamma = \text{band parameter}$$

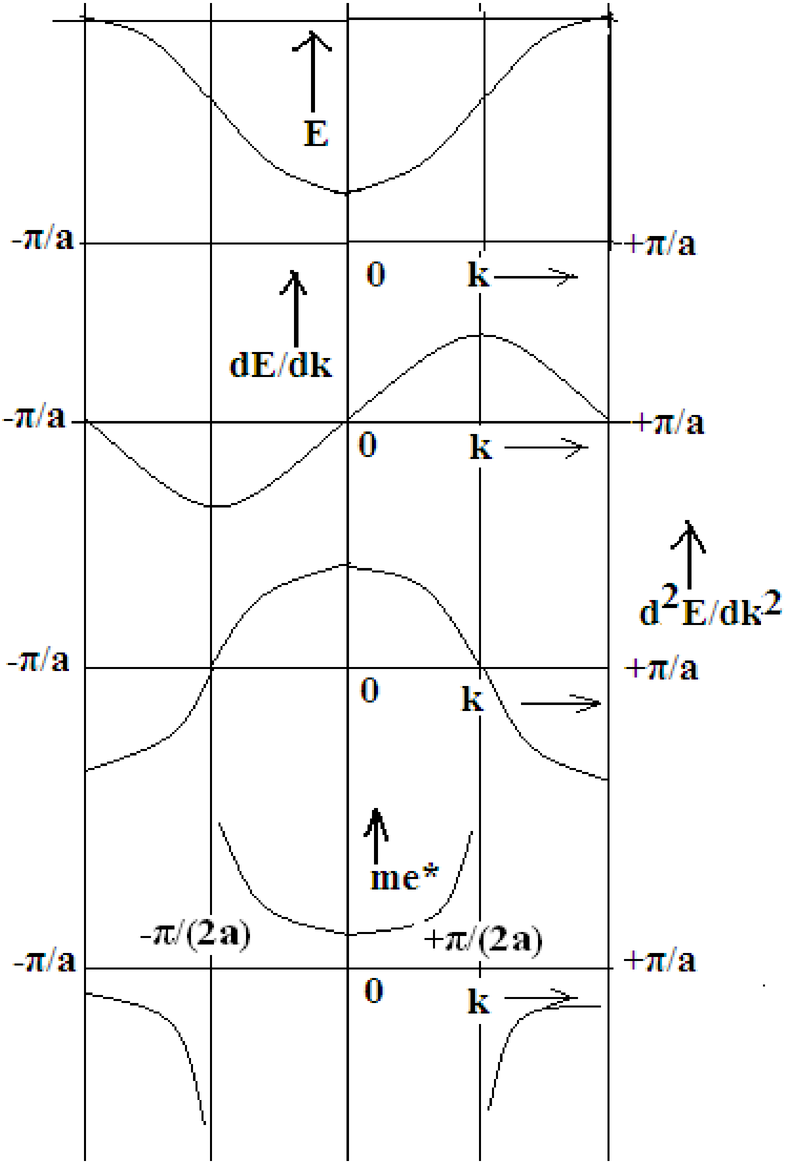


Figure 14. Reduced E-k , dE/dk-k, d 2 E/dk 2 -k and m*(effective mass of electron)-k diagram.π

From Figure 14 it is evident that Graphene is a zero-gap band semi-conductor or it is semi-metal. Because of zero bandgap , valence band and conduction band meet at K and K' Dirac points of sublattices A and B, as shown in Figure 5. The conducting electron is perpetually at the rock-bottom of the conduction conic band hence it has the lightest effective mass so much so that it is massless and Zhang et. al. (2005) have shown it to be 0.007m_e at Dirac Point. The velocity of a conducting electron with effective mass m_e^* subjected to an accelerating voltage V_{acc} is given as follows:

$$\frac{p^2}{2m_e^*} = qV_{acc} \qquad 8$$

$$v_e(velocity\ of\ electron) = \frac{1}{\sqrt{m_e^*}}\sqrt{2qV_{acc}} \qquad 9$$

Taking the effective mass of electron as 0.007m_e and accelerating voltage as 0.025V = kT where k= Boltzmann constant and T = 300Kelvin we obtain a velocity of 1.12×10^6m/s.

The velocity of conducting electron is a fraction of the velocity of light (~ c/300) hence electron in Graphene is a relativistic particle.

8.3.4.2.3. Energy of a relativistic paricle

Einstein equation of energy equivalent of a given mass m* is as follows:

$$E = m^* c^2 = rest\ mass\ energy + Kinetic\ Energy$$
$$= m_0 c^2 + \frac{p^2}{2m^*} \qquad 10$$

Where m_0 = rest mass and m* = effective mass moving with the relativistic velocity ‘v’.

The momentum is :

$$p = m^* v = \frac{m_0}{\sqrt{1 - \frac{v^2}{c^2}}} v \qquad 11$$

Multiply Eq.11 by ‘c’ and square the whole equation:

$$(pc)^2 = \frac{(m_0 v c)^2}{(1 - (\frac{v}{c})^2)} = \frac{((m_0 c)^2 [v^2])}{(1 - (\frac{v}{c})^2)} = \frac{((m_0 c)^2 c^2 [v^2/c^2])}{(1 - (\frac{v}{c})^2)}$$
$$= \frac{((m_0 c)^2 c^2 \left[\frac{v^2}{c^2} - 1 + 1\right])}{(1 - (\frac{v}{c})^2)} \qquad 12$$

Simplifying Eq.12 we get:

$$(pc)^2 = \frac{((m_0 c)^2 c^2 \left[\frac{v^2}{c^2} - 1\right])}{(1 - (\frac{v}{c})^2)} + \frac{((m_0 c)^2 c^2)}{(1 - (\frac{v}{c})^2)}$$
$$= -{m_0}^2 c^4 + (m^*)^2 c^4 \qquad 13$$

Substituting Eq 10 in Eq 13 we get:

$$(pc)^2 = -{m_0}^2 c^4 + E^2$$

14

Rearranging the terms in Eq.14 we get:

$$E^2 = (pc)^2 + (m_0c^2)^2$$

15

Equation 15 is the energy equation for relativistic particles with rest mass ‘m_0’.

8.3.4.2.4. Linear Energy Dispersion Spectra of conducting electron in Graphene.

We will use two equations from quantum mechanics namely:

Energy of electron where electron is a wave packet like a photon is given by the expression E (k) = hν = h(v_F/λ) = p.v_F16

Here v_F = c/300 = 10^6 m/s = Fermi Velocity = the velocity with which zero rest mass electron zips around the 2-D crystalline lattice of graphene.

Using de Broglie Equation p = h/λ= ±ħ□k□ in Eq.16 we get:

E (k) = ± ħv_F.□k□17

where k = wave vector and Fermi Energy Level is the Dirac point where E =0.

Linear Energy Dispersion Spectra gives rise to Energy Cone as shown in Figure 5.

The above linear energy dispersion gives rise to Dirac quasi-particles. Dirac-like equation is a consequence of graphene crystal structure which consists of two equivalent carbon sub-lattices A and B. Quantum Mechanical hopping between the sub-lattices leads to the formation of two cone-like energy bands and their intersection, the Dirac Points K and K', near the edges of Brillouin Zone yields the conical energy spectrum as shown in Figure 5. As a result , quasi-particle in graphene exhibit linear dispersion relation given by Eq.17. The particles in conventional metals and semi-conductor have parabolic (free-electron-like) dispersion relation. Because of this difference in energy dispersion relations the quasiparticles in graphene behave very differently from the particles in 3-D solids.

The impact of the topological effects on graphene’s electronic properties can be elegantly described by the Atiyah-Singer Index Theorem. A direct consequence of the index theorem is charge fractionalization which is known as fractional quantum effect. The charge fractionalization gives rise to the possibility of realizing grapheme based Anyons that unlike Bosons or Fermions exhibit fractional Statistics. Anyons are a strong candidate for performing error free quantum information processing. [Pachos (2009)].

8.3.4.2.5. Electrons and Holes in Graphene behave like Electrons-Positrons of Quantum Electro Dynamic (QED)Systems- The Origin of Charge Conjugate Symmetry.

[Katsnelson, Novoselov & Geim (2006)]

Electrons and holes in classical Solid State Devices or in Quantum 3D devices behave as independent entities, they are Schrodinger particles and they have separate Schrodinger descriptions. But this is not the case for electron-holes in 2D grapheme.

As seen in Figure 15, filled energy states at the rock-bottom of the conduction cone gives rise to conducting electrons with positive energy, negative charge and zero rest-mass. This contributes to the electronic conduction in graphene.

As a duality of this electronic conduction, the empty energy states at the apex of the valence cone corresponds to negative energy, positive charge and this also has zero rest mass. These empty energy states are holes and are analgous to positrons in Quantum Electro Dynamics(QED). This is a direct consequence of Time-Reversal Invariance.

Just as in QED, we have one spinor function to describe electron-positron pair. Analogously we have the same spinor function to describe electron-hole pair in graphene except that the real spin of electron-positron is replaced by the pseudo-spin (σ) of electron-hole pair. Mapping of sublattices degree of freedom to pseudospin is best represented by Dirac Equation.

In nutshell the charge-conjugation-symmetry found in QED is exhibited in graphene also.

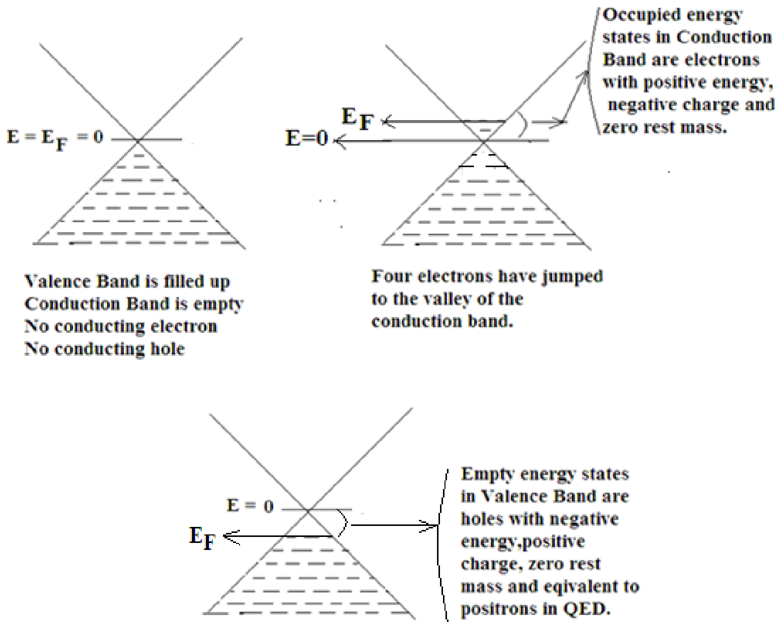
8.3.4.2.6. Chirality of Electrons and Holes in Graphene.

Figure 16 describes the hexagonal unit cell. As shown in Figure 16, the chiral vector determining the structure of CNT is given by ‘L’ and its length gives the circumference of CNT. The chiral angle is represented by ‘η’. If $\eta = 0$ then CNT has zigzag periphery as shown in Figure 4.b and if $\eta = \pi/6$ then CNT has arm-chair periphery as shown in Figure 4.c. [Ando (2009)].

Figure 17 describes the honeycomb structure consisting of two interpenetrating triangular lattices(“A” and “B” sublattices) with one lattice point of each type per unit cell. 2-D inversion (i.e. rotation by π) interchanges the two sublattices [Haldane (1988)].

In Figure 17, Wigner-Seitz Unit Cell is centered on the star point ‘*’ of 6-fold symmetry and is bounded by the hexagon of nearest neighbor bond. Arrows on the second-neighbour bonds mark the direction of positive hopping with broken time reversal invariance.

In Figure 18, we have shown the tunneling of electron through a potential barrier V_0 high and D wide. *Projection of pseudospin 'σ' on the direction of motion is chirality.* Hence chirality is positive for electron and negative for holes. Chirality shows the additional built-in symmetry between electron and hole parts in graphene spectrum and it is analogous but not identical to chirality in 3-D materials. Thus we have two built-in symmetries in graphene: *charge conjugation symmetry and Chirality symmetry*.



Occupied energy states in Conduction Band are electrons with positive energy, negative charge and zero rest mass.

Valence Band is filled up
Conduction Band is empty
No conducting electron
No conducting hole

Four electrons have jumped to the valley of the conduction band.

Empty energy states in Valence Band are holes with negative energy,positive charge, zero rest mass and eqivalent to positrons in QED.

The hexagonal lattice symmetry and because of the two sublattices A and B contribution symmetry, electrons and holes have conjugation charge symmetry just as electron and positrons have conjugation charge symmetry in QED. Analogously the same spinor wave function which describes electron and positron in QED can also describe electrons and holes in Graphene except that the real spin of electron-positron will be replaced by psuedo-spin (σ) of the electron-hole pair.
3-D semi-conductors —— electrons and holes have different effective mass and different Schrodinger functons describing the two particles.

Figure 15.The origin of Charge–conjugation-

symmetry of electron and holes in Graphene and this results in spinor function description of electrons and holes in Graphene just as we have spinor function description of electron-positron in QED.

In essence the chiral character of the electronic charge in graphene is due to pseudospin associated with the two triangular sub-lattices that form the honeycomb lattice.

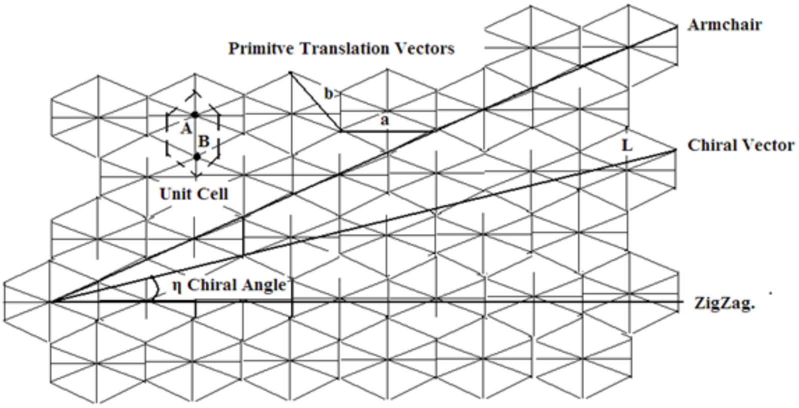


Figure 16.The honeycomb 2D lattice of Graphene. The hexagonal unit cell contains two atoms (A and B). The chiral vector determining the structure of CNT is given by 'L' and its length gives the circumference of CNT. The chiral angle is represented by 'η'. If $\eta = 0$ then CNT is zigzag and if $\eta = \pi/6$ then CNT is armchair.

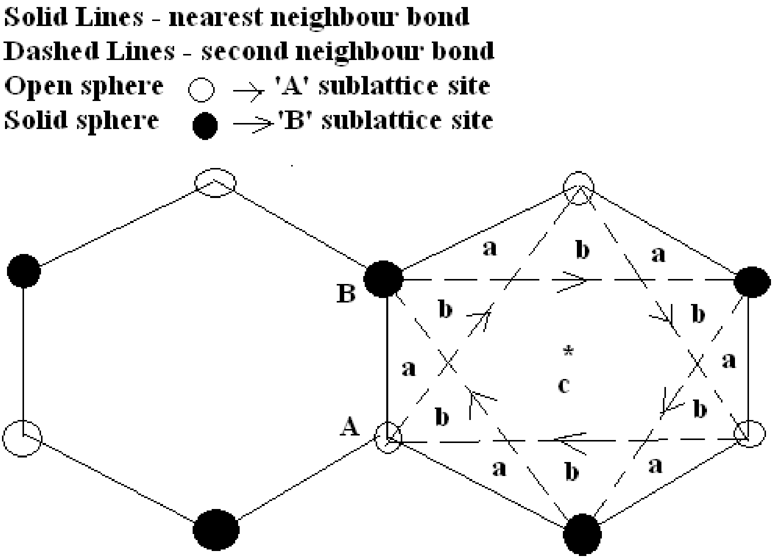

Solid Lines - nearest neighbour bond
Dashed Lines - second neighbour bond
Open sphere ○ → 'A' sublattice site
Solid sphere ● → 'B' sublattice site
B
A
a
b
a
b
b
a
a
*
c
b
b
a
a
b
a

Figure 17. 2-D grapheme has the honeycomb structure consisting of two interpenetrating triangular lattices (“A” and “B” sublattices) with one lattice point of each type per unit cell. 2-D inversion (i.e. rotation by π) interchanges the two sublattices. [Haldane (1988)]

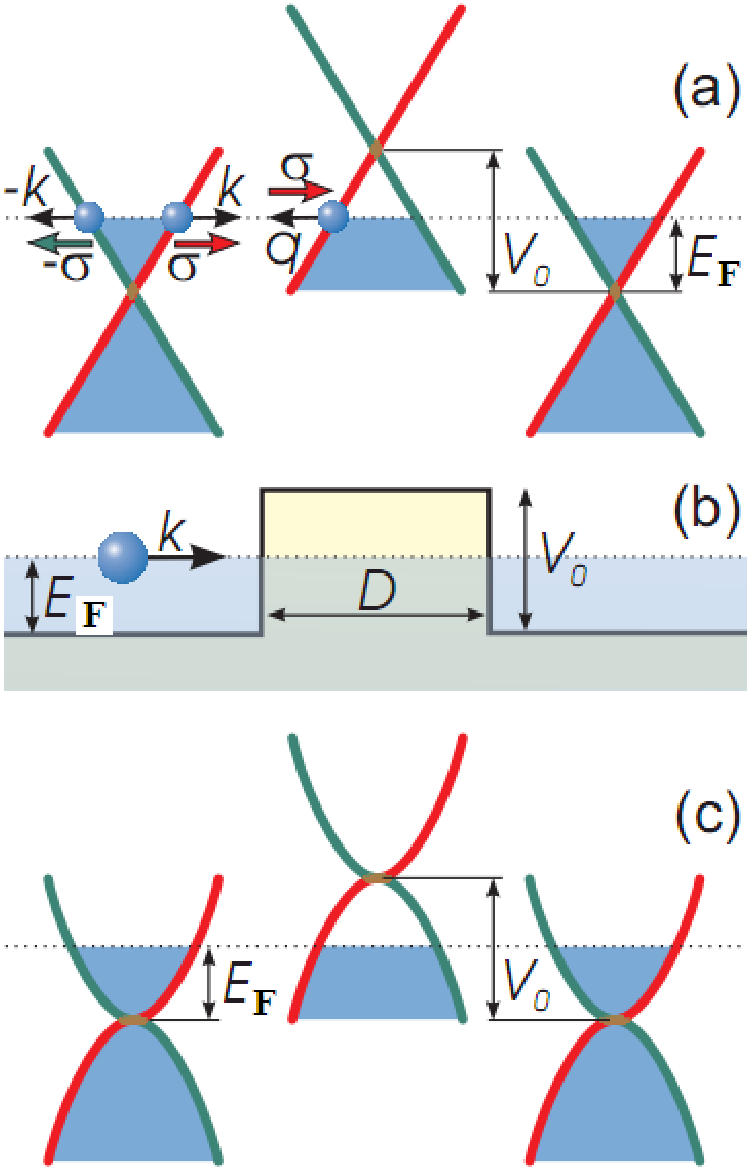


Figure 18. Tunneling through a potential barrier in graphene. (a)-Schematic diagrams of the spectrum of quasiparticles in single layer graphene. The spectrum is linear at low Fermi

energies (<1 eV). The red and green curves emphasize the origin of the linear spectrum, which is the crossing between the energy bands associated with crystal sub-lattices A and B. The three diagrams illustrate schematically the positions of the Fermi energy E F across the potential barrier of height V 0 and width D shown in (b). The Fermi level (dotted lines) lies in the conduction band outside the barrier and the valence band inside it. The blue filling indicates occupied states. The pseudospin denoted by vector σ is parallel (antiparallel) to the direction of motion of electrons(holes),which also means that σ keeps a fixed direction along the red and green branches of the electronic spectrum.

(c)-Low energy spectrum for quasiparticles in bilayer graphene. The spectrum is isotropic and, despite its parabolicity, also originates from the intersection of energy bands formed by equivalent sublattices,which ensures charge conjugation, similar to the case of single-layer graphene.

8.3.4.2.7. Schrodinger particle in contact and in bulk semi-conductor/graphite and Dirac quasi-particle in graphene.

In 3-D metallic contacts and in 3-D graphite/ semiconductor, electron has a finite rest mass and negative charge of ‘e’ Coulombs and is described by Schrodinger Equation as follows:

Schrodinger Equation:

$$H\Psi = E\Psi \qquad 18.a$$

$$\text{Where H} = \text{Hamiltonian Operator} = \frac{\mathrm{p}^2}{2\mathrm{m}^*} + \mathrm{V(x)}$$

$$p = \frac{\hbar}{i}\frac{\partial}{\partial x} \text{ and } H = i\hbar\frac{\partial}{\partial t}$$

Letting the Hamiltonian operate on matter wave function we get the final Schrodinger Equation as follows:

$$i\hbar\frac{\partial}{\partial t}\Psi(x,t) = \frac{-\hbar^2}{2\mathrm{m}^*}\frac{\partial^2}{\partial x^2}\Psi(x,t) + V(x)\Psi(x,t) \qquad 18.b$$

By separation of variables, the space dependent part and the time dependent partr are separated and then the second order differential equation is solved. The total solution is as follows:

$$\Psi\ (probability\ amplitude) = A_1\,[\exp(ik_1x) + B_1\exp(-ik_1x)]exp(i\omega t) \qquad 18.c$$

$$\Psi^*\Psi = |\Psi|^2 = probability\ density = probability\ of\ finding\ electron\ per\ unit\ volume.$$

Here the first term is Backward wave-function and the second term is the Forward wave-function and the wave vector :

$$k_1 = \frac{\sqrt{2m^*(E+eV)}}{\hbar} \qquad 19$$

Where E = total energy = kinetic energy [p2/(2m*)] + potential energy(-eV).

A1 and B1 are two arbitrary constants incorporated in the complementary solution of 2nd order ordinary linear differential equation. These two are determined from two boundary conditions of the real physical situation.

In graphene, we have to take the relativistic version of Schrodinger Equation. Relativity treats Space and Time as an unified space-time so relativistic generalization requires that space and time derivatives must enter symmetrically as they do in Maxwell Equations of Electro-Magnetic Field Theory. In relativity, momentum and energy are the space and time of 4-D space-time vector (x, y, z, t). and they are related by the relativistic ally invariant relation derived in Section 4.4.2.3, Equation 15:

$$\frac{E^2}{c^2} - p^2 = m_0^2 c^2 \qquad 20.a.$$

This ‘Equation 20.a.’ says that the length of the vector is proportional to the invariant rest mass m0. Substituting the operator equivalents we get an equation describing the propogation of waves constructed from relativistic invariant objects, the Klein-Gordon Equation. In this relativistic formulation, Ψ wave function is a relativistic scaler: a complex number which has the same numerical value in all frames of reference. In context of Quantum Field Theory, the □Ψ□2 corresponds to charge density and it can be positive and it can be negative. In Schrodinger formulation it is always probability density.

Finding a relativistic field equation with first order derivatives required a more elaborate construction. The original form of Dirac Equation is:

$$(c\boldsymbol{\alpha}.\boldsymbol{p}+\boldsymbol{\beta} mc^2)\boldsymbol{\Psi} = i\hbar\frac{\partial}{\partial t}\Psi(x,t) \qquad 20.b$$

$\alpha = \alpha_1+\alpha_2+\alpha_3$.

Here $\alpha_1,\alpha_2,\alpha_3$ and β are 4×4 matrices. They are all Hermitians. Hence Dirac Hamiltonian is Hermitian.

Electron in 2-D graphene lattice behaves like Dirac quasi-particle which is a neutrino with electronic charge ‘e’. It is a relativistic particle with zero rest-mass and described by Dirac Spinor Functions:

$$\Psi = \begin{bmatrix}\Psi_1\\ \Psi_2\end{bmatrix} = \begin{bmatrix} A_2\exp(ik_2x)+B_2\exp(-ik_2x) \\ S_2[A_2\exp(ik_2x)+B_2\exp(-ik_2x)]\end{bmatrix} \qquad 20.c$$

Here the wave vector k_2 is given by the wave vector formulation of a relativistic quasi-particle of zero-rest mass as given Eq.4. The more generalized wave vector formulation of k2 is given as follows:

$$k_2 = \frac{(E - U_g + \frac{eV}{2})}{\hbar v_F} \qquad 21$$

In graphene devices, semi-conductor physics has to be redone with Dirac Equations. We have to explore issues in Relativistic Quantum Mechanics which are otherwise inaccessible to experiment. In graphene we have to exploit the mathematical beauty of Dirac Equation to the best in developing novel devices and novel Integrated Circuits based on graphene.

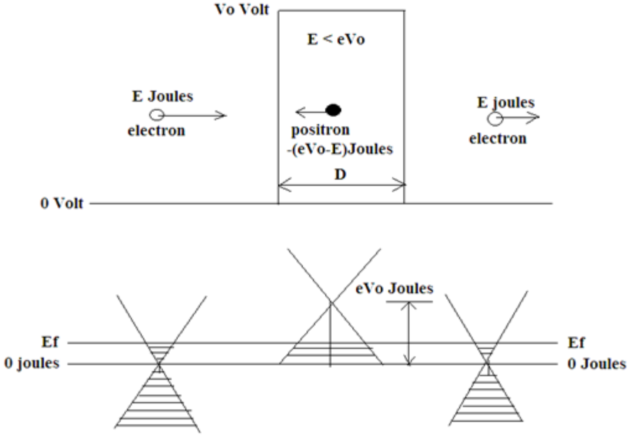
The Integrated Circuit Technology of Future Generations-Part 5_QED properties continued
The Integrated Circuit Technology of Future Generations-Part 5 describes Klein Paradox,Anolmous Integer Quantum Hall Effect,Aityah-Singer Index Theorem,near perfect crystal and opening the band-gap.

8.3.4. I.C.Technology of Future Generations_Part 5_Quantum Electrodynamical Properties of Graphene continued.

8.3.4.3. Klein Paradox.

[Katsnelson,M.;Novoselov,K.,& Geim,A.K., "Chiral tunneling and Klein Paradox in Graphene", *Nature Physics*, **2,** 620-625, (2006);]

Unimpeded penetration of relativistic particles through high and wide potential barriers is known as Klein Paradox and is a consequence of Quantum Electrodynamics (QED).



Fermi Energy Level is in Conduction Band outside the potential barrier therefore there are filled up energy states in Conduction Band. These filled up states correspond to Dirac Electrons.
Fermi Energy Level is in Valence Band inside the potential Barrier therefore there are empty energy states at the top of the Valence Band. These empty energy states at negative energy level -(eVo-E) correspond to Dirac Positrons.

Figure 19. The creation of Dirac Electron-Positron at the right interface of the Potential barrier and annihilation of Dirac Electron-Positron at the left interface of the potential barrier.

As seen in Figure 19 for every incident Dirac Electron at the left of the potential energy barrier of height eV_0 Joules there is a conjugate Dirac Positron in the Potential Energy Barrier and there is a corresponding transmitted Dirac Electron on the right of the barrier.

At the right interface, Dirac electron-positron is created. Dirac Electron moves towards right as the transmitted electron and Dirac Positron moves leftward towards the left interface of the barrier.

At the left interface, incident Dirac Electron and leftward moving Dirac Positron annihilate each other.

This creation-annihilation of electron-positron in QED system is equivalent to generation-recombination of electrons-holes in 3-D semi-conductors.

Thus matching between Dirac electron and Dirac positron spinor wave functions across the barrier leads to the high probability tunneling described by Klien Paradox. This Transmission probability ‘T’ depends only weakly on the barrier height. It approximates to perfect transparency for very high barriers. In non-relativistic tunneling ‘T’ exponentially decays with increasing height.

8.3.4.3.1. Klein Paradox reformulated for single layer 2D Graphene

(Setting up the spinor functions outside and inside the Potential Barrier and match the reflection and transmission cofficients).

Refer to Figure 19.

$V(x) = V_0$ for $0 < x < D$;

$V(x) = 0$ otherwise.

We assume abrupt or step-like barrier edges. This assumption is valid if the Fermi-Wavelength λ of the quasi-particles is much larger than the characteristic width of the edge smearing which in turn should be larger than the lattice constant ‘a’ (to disallow Umklapp scattering between different valleys in graphene).

Dirac Electron Fermions are massless hence there is no formal requirement for the minimal electric field (ε) to form positron like states under the barrier. So for graphene experiments we routinely take the opposing field of the potential barrier to be $\varepsilon = 10^5$V/cm . A field of this order is sufficient to create a well-defined potential barrier in realistic graphenes samples with disorder.

For experimental verification of Klein Paradox in QED system, a potential drop of (m_ec^2/e)Volt is required over the Compton Length ($\hbar/(m_ec)$) meter within the barrier. This works out to be 10^{18}V/m = 10^{16}V/cm. This sort of barrier can be encountered only in extra –ordinary situation such as positron production in super-heavy nuclei with charge $Z \geq 170$ [Greiner,W.,Mueller,B.,andRafelski,J.(1994), Grib,A.A.,Mamayev,S.G.,andMostepanenko,V.M.,(1994)] or in the evaporation of Black holes through generation of particle-anti particle generation at the event horizon [Page, D. N.(2005)].

These same very relativistic quantum tunneling described by Klein Paradox and other related QED phenomena can be tested in graphene test bed.

We assume that the incident Dirac electron wave travels at an angle ‘φ’ with respect to x-axis as shown in Figure 20.

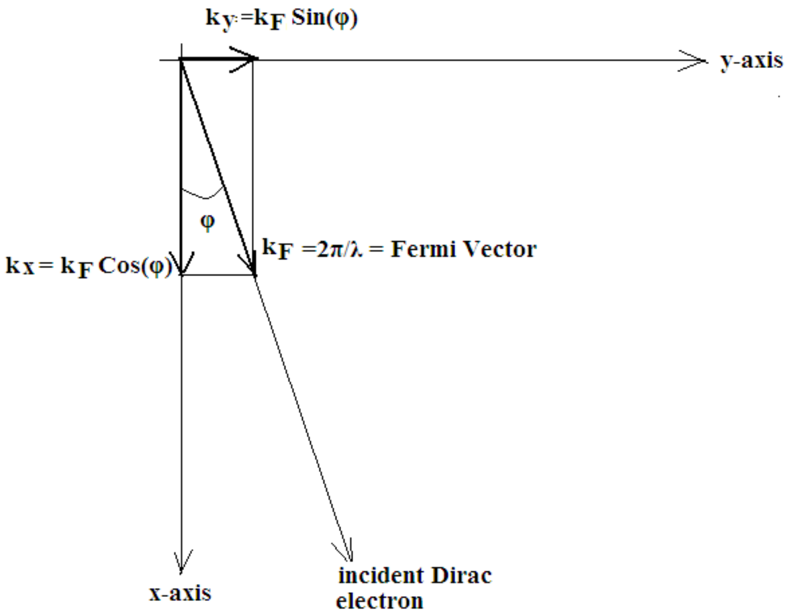


Figure 20. The incidence angle ‘φ’, its propogation Fermi-vector and its projections on the x-axis and y-axis.

We write the Dirac spinor functions Ψ1 and Ψ2 for the Hamiltonian H = E(Kinetic energy) + eV(potential energy):

$$\Psi 1(x, y) = \left[e^{ik_x x} + re^{-ik_x x}\right] \times e^{ik_y y} \quad for\, x < 0;$$

$$= \left[ae^{iq_x x} + be^{-iq_x x}\right] \times e^{ik_y y} \quad for\ 0 < x < D;$$

$$= te^{ik_x x + ik_y y} \qquad for\ x > D;$$

$$\Psi 2(x, y) = S\left[e^{ik_x x + i\varphi} + re^{-ik_x x - i\varphi}\right] \times e^{ik_y y} \quad for\ x < 0;$$

$$= S'\left[ae^{iq_x x + i\theta} + be^{-iq_x x - i\theta}\right] \times e^{ik_y y} \quad for\ 0 < x < D;$$

$$= Ste^{ik_x x + ik_y y + i\varphi} \qquad for\ x > D;$$

Where

$q_x = \sqrt{\frac{(E - eV_0)^2}{\hbar^2 v_F^2} - k_y^2}$;$refraction\ angle = \theta = Tan^{-1}\left(\frac{k_y}{q_x}\right)$; $S = Sign(E)$; $S' = Sign(E - eV_0)$;

Satisfying the continuity of the wave function by matching the coefficients a, b, r and t we get the reflection coefficient ‘r’ as follows:

$$r = 2ie^{i\varphi} Sin(q_x D) \times \frac{Sin\varphi - SS'Sin\theta}{SS'[e^{-iq_x D} Cos(\varphi + \theta) + e^{iq_x D} Cos(\varphi\theta) - 2iSin(q_x D)}]$$

Transmission = T = □t□2 =1-□r□2

In the limit of high barrier where □eV0□ > > □E□ we get:

$$T = \frac{Cos^2\varphi}{1 - Cos^2 q_x D \times Sin^2\varphi} \qquad 22$$

Under resonance conditions when $q_x D = \pi N$ where $N = 0, \pm 1, \ldots$.

$T = 1$;

For angles close to Normal Incidence, T is always equal to Unity.

The matching between directions of psuedospin ‘σ’ for quasi-particles inside and outside the barrier results in perfect tunneling.

The Fermi wavelength is given as follows:

$$\lambda = \frac{v_F}{\nu} = \frac{v_F h}{E_F}$$

For Fermi Energy 80meV and Fermi-velocity = 10^6m/s, λ = 50nm.

In non-relativistic Quantum devices such as Zener Diodes and Esaki Diodes the transmission coefficient is given as:

$$T = Exp\left[\frac{-2D\sqrt{(2m_e(eV_0 - E))}}{\hbar}\right] \tag{23}$$

So as the height of the potential barrier increases, transmission falls exponentially.

Figure 21 gives the transmission pattern across the 'D' unit wide and "V0" volt high potential barrier with respect to incident angle 'φ'.

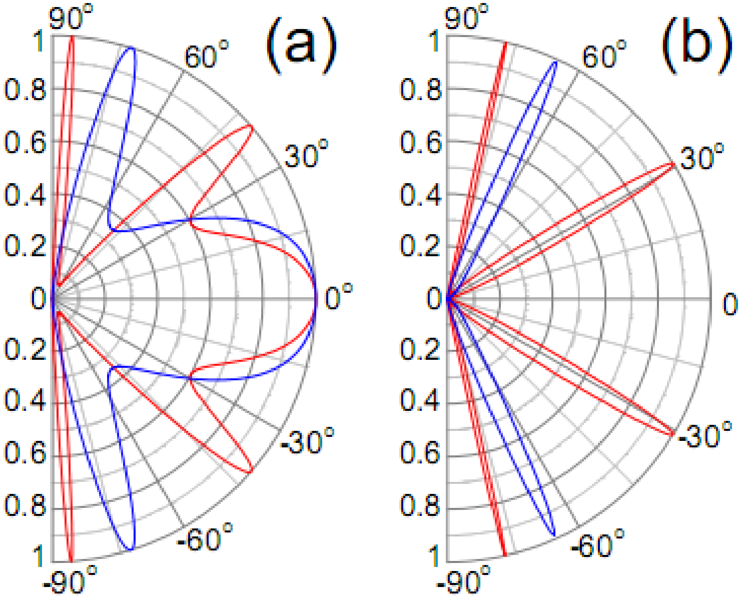


Figure 21.Klein-like quantum tunneling in graphene systems.Transmission probability T through a 100-nm-wide barrier as a function of the incident angle for (a)single-and(b)bi-layergraphene. The electron concentration n outside the barrier is chosen 0.5 × 10 12 cm − 2 for all cases.Inside the barrier,hole concentrations p are1 × 10 12 and 3 × 10 12 cm − 2 for red and blue curves, respectively (such concentrations are most typical in experiments with graphene).This corresponds to the Fermi energy E of incident electrons ≈ 80 and 17 meV

for single-and bi-layer graphene, respectively,and $\lambda \approx$ 50nm.The barrier heights V 0 are(a)200and(b)50meV(red curves) and (a)285 and (b) 100meV(blue curves).

8.3.4.4. Anolmous Integer Quantum Hall Effect.

In 2-D electron gas in graphene when a magnetic field is applied perpendicular to the sheet, electron and hole orbits are created. These orbits are created at quantized values of Energy. These are known as Landau Levels of energy.

There were two groups of researchers working on this ‘Anolmous Quantum Hall Effect” after the discovery of graphene.Phillips Kim was leading a group at Columbia University in USA and A.K. Geim’s was leading a second group at University of Manchester, England, and at the Institute for Microelectronics Technologyin Chernogolovka, Russia. Both group discovered that when graphene is sufficiently cold and exposed to high magnetic field then the sheet exhibits a distinctive Quantum Hall Effect.

Graphene exhibits exotic fractional quantum Hall States even at room temperature [Novoscelov et.al. (2005),Zhang et.al.(2005), Zheng et.al.(2002)] extending the previous temperature of observation by a factor 10 . Integer Quantum Hall Effect is anolmous in that it occurs as half integer filling factors. Hence we call it Half-Integer Quantum Hall Effect given as below[Zhang et.al.(2005)]:

$$R_{XY}^{-1} = Hall\ conductivity = \sigma_{Hall} = \frac{g_s e^2}{h}\left(N + \frac{1}{2}\right)\ where\ N = Landau\ Level(LL)\ \ or\ Landau\ index = \frac{no.of\ electrons\ per\ cm^2}{no.of\ magnetic\ flux\ lines\ per\ cm^2}\ \ (at\ 4K, 14T) \qquad 21$$

In Equation 21, $g_s = 4$ which is the four fold degeneracy due to spin and sublattices. Landau Index = +1,+2… for electrons and -1, -2,…. for holes and Hall Conductivity occurs for N=0 also. As soon as we go to bi-layer of graphene, the band symmetry is altered and ½-integer phase shift is removed [Wilson (2006)].

The half-integer shift is a subtle consequence of graphene's symmetric topology which constrains the electrons to obey Dirac Relativistic Equation. In Section 3.4.2.6 we had introduced the concept of pseudo spin and chirality. As electron orbits one cycle around the magnetic flux, its pseudospin rotates introducing 180° phase shift in electron wave function. This half-wavelength shift is Berry Phase or Geometric Phase discussed in Section……. The introduction of Berry Phase changes the pattern of allowed energy levels and quantized steps in Hall Conductivity by just the observed amount.

Unique properties of Dirac quasi-particles in 2-D graphene such as 2-fold degeneracy for spin and 2-fold degeneracy for two sub-lattices, electron-hole degeneracy and vanishing carrier mass near the point of charge neutrality, know as Dirac point, gives rise to distinctive half-integer Quantum Hall Effect which was already theoretically predicted [Zheng et. al.(2002), Gusynin et.al.(2005), Peres et.al.(2005)]. Using(2space dimensions + 1time dimension) dimensional QED, Zheng et.al.(2002) had proved that Hall conductivity in graphene displays Quantum Hall Effect when E_F is in low lying Landau Levels and scattering is weak. When E_F is far away from zero energy level then Hall Conductivity obtained by Self-Consistent Born Approximation approaches Boltzmann Conductivity.

Zhang et. al. (2005) have experimentally measured R_{XX} (magnatoresistance) and R_{XY} (Hall Resistance) in a very good quality graphene samples and discovered half-integer shifted Quantum Hall Effect as shown in Figure 22.

Conventional 2D devices exhibit Integer QHE. The half-integer shift observed in graphene is a result of the topologically exceptional electronic band structure of graphene. Time reversal invariance imparts charge conjugation symmetry to electron-hole. They both act as electron-positron pair with zero rest mass. (2 + 1) dimensional QED analysis gives Landau Level Energy [Zhang et. al.(2005)]. The relativistic Landau Level and charge conjugation symmetry have resulted into half-integer shifted QHE

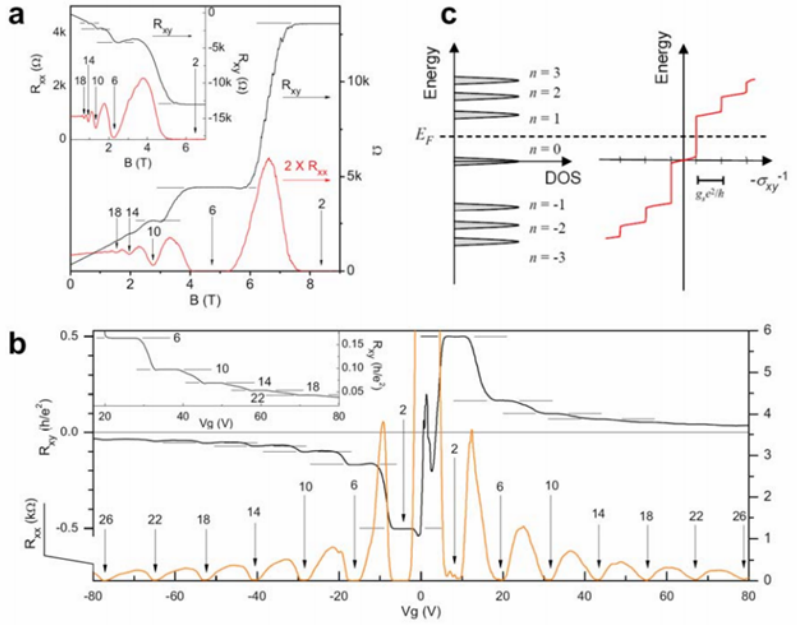


Figure 22. Quantized magnetoresistance and Hall resistance of a graphene device. [Zhang.et.al.(2005)]

a, Hall resistance (black) and magnetoresistance (red) measured in the device at T = 30 mK and Vg = 15 V. The vertical arrows and the numbers on them indicate the values of B and the corresponding filling factor ν of the quantum Hall states. The horizontal lines correspond to $(h/e^2)\times\nu$ values. The QHE in the electron gas is demonstrated by at least two quantized plateaus in Rxy with vanishing Rxx in the corresponding magnetic field regime. The inset shows the QHE for a hole gas at Vg = -4 V, measured at 1.6 K. The quantized plateau for

filling factor $\nu = 2$ is well-defined and the second and the third plateau with $\nu = 6$ and 10 are also resolved.

b, The Hall resistance (black) and magnetoresistance (orange) as a function of gate voltage at fixed magnetic field B = 9 T, measured at 1.6 K. The same convention as in (a) is used here. The upper inset shows a detailed view of high filling factor plateaus measured at 30 mK.

c, A schematic diagram of the Landau level(LL) density of states (DOS) and corresponding quantum Hall conductance (σ_{xy}) as a function of energy. Note that in the quantum Hall States,$\sigma_{xy} = -R_{xy}^{-1}$. The LL index n is shown next to the DOS peak. In our experiment, the Fermi energy E_F can be adjusted by the gate voltage, and R_{xy}^{-1} changes by an amount of (g_s) ×(e^2/h) as E_F crosses a LL.

As seen in Figure 22 , magneto-resistance is always positive hence its plot with respect to gate voltage is an even function and Hall Resistance is negative for holes and positive for electrons hence it is odd function with respect to gate voltage. Since Hall Resistance is discretized or quantiuzed hence clearly resolved plateau occur at $N = 0, \pm 2, \pm 6, \pm 14$. At the points where Hall Conductivity $R_{XY}(B)$ peak occurs just at those points magneto-resistance R_{XX}

vanishes. This is the hall-mark of QHE.

½ Integer QHE is altered to integer QHE even if one additional sheet is added to the system as shown in Figure 23.

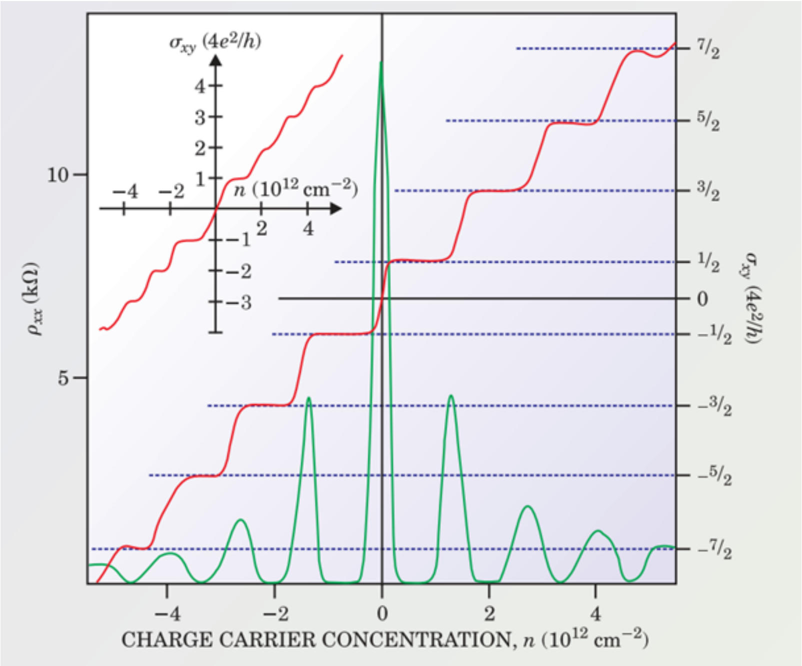


Figure 23. Quantum Hall effect in graphene. The hallmark of the integer QHE

is that the longitudinal resistivity ρ xx (shown in green) of a two-dimensional

electron gas in a high magnetic field vanishes at precisely those discrete val-

ues where the transverse conductivity σ xy (red)

is quantized in units of e 2 /h,

where h is Planck’s constant and e the charge on an electron or hole. An extra

factor of 4 accounts for the spin and band-structure degeneracy in graphene.

At 4 K and 14 T, σ xy increases in quantized steps as researchers sweep the

gate voltage applied to the crystal to vary the electron density n. But those

steps come at integer values shifted by one-half of a unit because electrons

mimic the relativistic behavior of Dirac fermions. Adding even one additional

sheet to the system alters the band symmetry and removes the 1/2-integer

phase shift, as the inset plot shows. [Adapted from Wilson (2006)] .

The two sublattice symmetry gives the quantum phase to spin-1/2 pseudo-spinor function describing the Dirac quasi-particles. This quantum phase is known as non-trivial Berry Phase or Geometric Quantum Phase already described in Section 3.1. This Berry Phase is implicit in half-integer shifted quantization rules of QHE. This Berry Phase can further be probed in the magnetic field region where semi-classical magneto –oscillation is manifested [Sharapov et.al. (2004); Lukyanchuk and Kopelevich (2004)]. This is known as Shubnikov de Hass (SdH) Oscillation. SdH oscillation is described by Equation 22 and illustrated in Figure 24.

$$R_{XX} = R(B,T)Cos\left[2\pi\left(\frac{B_F}{B}+\frac{1}{2}+\beta\right)\right] \qquad 22$$

Where R(B,T) = SdH Oscillation Amplitude; B_F = the frequency of oscillation in 1/B;

And β = associated Berry Phase which lies between ZERO and UNITY. β=0 and β=1 are trivial cases and β=1/2 implies Dirac Particles.

From the temperature dependence of well developed SdH Oscilations at low magnetic field using standard SdH formalism we can extract the effective carrier mass m*. The effective mass comes to be $0.007m_e$ at the Dirac Point and it is $0.04m_e$ at a gate voltage corresponding to carrier density of $5\times10^{12}/cm^2$.

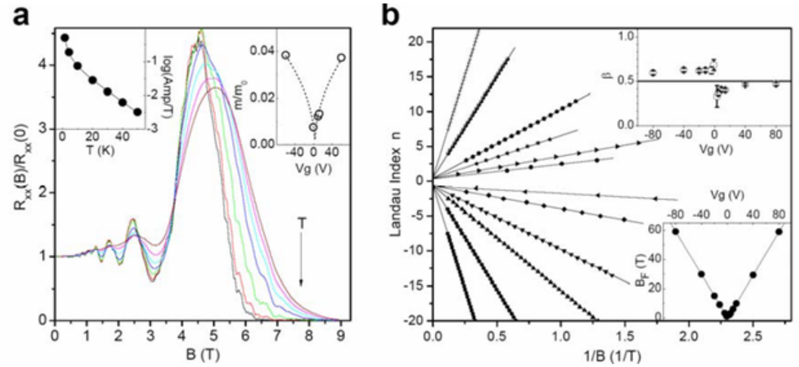


Figure 24. Temperature dependence and gate voltage dependence of the Shubnikov de Haas oscillations in graphene. a, Temperature dependence of the SdH oscillations at Vg = -2.5 V. Each curve represents Rxx (B) normalized to Rxx (0) at a fixed temperature. The curves are in order of decreasing temperature starting from the top as indicated by the vertical arrow. The corresponding temperatures are listed in the left inset. The left inset represents the SdH oscillation amplitude divided by temperature measured at a fixed magnetic field. The standard SdH fit yields the effective mass. The right inset is a plot of the effective mass obtained at different gate voltages. The broken line is a fit to the single parameter model described in the text, which yields vF = 1.1x106 m/s, in reasonable agreement with the literature values. b, A fan diagram for SdH oscillations at different gate voltages. The location of 1/ B for the n th minimum (maximum) of Rxx counting from B = B F is plotted against n (n +1/2). The lines

correspond to a linear fit, where the slope (lower inset) indicates *B* F and the *n* -axis intercept (upper inset)

provides a direct probe of Berry's phase in the magneto-oscillation in graphene.

Non-zero Berry's Phase (or geometric Quantum Phase) had also been predicted because of the exceptional topology of graphene band structure [Ando et.al.(1998), Mikitik & Sharlai (1999)]. With the fabrication of Graphene [Novoselov et.al.(2004); Berger et.al. (2004); Zhang et.al. (2005a); Zhang et.al. (2005b); Bunch et.al.(2005)] the experimental verification of these exotic features have become a reality.

8.3.4.5. Aityah-Singer Index Theorem [Pachos (2009)]

The impact of the topological effects on grapheme electronic properties can be elegantly described by the Aityah-Singer Index Theorem. This predicts the number of permissible quantum states of the electron at the apex of the Dirac Cone. A direct consequence of the Index Theorem is charge fractionalization that is known as fractional quantum effect. The charge fractionalization gives rise to the possibility of realizing grapheme based Anyons that unlike Bosons and Fermions exhibit fractional statistics. Anyons are strong candidate for performing error free quantum information processing.

8.3.4.6. Graphene is near perfect single crystal free of structural defects but with many caveats.

Solid State Device Designer’s wish list consists a material which has no crystalline defects and precisely controlled impurities and no processing damages. Graphene is expected to meet these criterian but not quite. The underlying substrate has an influence on the structure and hence on the transport properties of the overlying graphene [Morozov et. al.(2006) 9]. The lithographic technique used for nano-patterning introduces photo-resist residues on the devices. These residues behave as adsorbates and are expected to effect the transport properties of graphene [Schedin et.al.(2007) 7]. Local charge disdorder [Huang et.al.(2006) 10] influence the transport of electrons. Atomic structure and the dangling bonds at the periphery of the graphene [Nimmi et.al.(2006) 11] and atomic scale defect [Wehling et.al.(2007) 12] play an important role in the electronic properties of graphene.

Ishigami et.al.(2007) have resolved atomic structures of oxide-supported graphene based electronic devices using SEM/AFM/STM techniques. Real space images of graphene atomic lattice was taken and its thickness characterization and nano scale corrugation characterization of a clean sample was done. It was found that graphene partially conformed to the underlayer SiO_2 because graphene haver finite intrinsic stiffness. Resist residues are ubiquitious in devices using lithographic nano-patterning. They have to be fully removed if the intrinsic structural properties of graphene are to be revealed and taken advantage of. This study enabled to assess the impact of atomic scale defects and adsorbates on nano-scale transport properties.

The corrugation of the graphene due to underelying silicon-dioxide substrate lowers the mobility of electrons [Morozov et.al.(2006)] and suppresses weak localization in graphene based devices[ibid]. Trapped Oxide charges of Silicon Dioxode also effect the transport properties of graphene [13]. The chemical adsorbates caused by Acrylic Lithography introduce unknown and uncontrollable perturbation in graphene.

Heat treatment at 400°C in an atmosphere of Argon/ Hydrogen atmosphere completely removes the resist residue. This helps uncover the intrinsic properties of graphene. Atomic-Resolution images of the graphene lattice show triangular and hexagonal lattice pattern in close proximity and this reduces the mobility of carriers.

Hass et.al. (2006) have studied the structural coherent domain size, surface roughness from sublimation, mobility of electron and long range order in graphene grown on C-faced SiC substrate(0001bar) and in graphene grown on Si-faced SiC substrate(0001). On all these counts epitaxially grown graphene(EG) on C-faced SiC is far superior. Mean coherent structural domain size is three times larger, surface roughness due to sublimation in SiC is suppressed by an order of magnitude, mobility is much larger (in EG grown on C-faced substrate μ= 10,000 to 27,000cm^2/(V-s) and in Si-faced sample μ= 1,000 cm^2/(V-s) and long range order exceeds 1 micron in EG on C-faced SiC.

This study proves that substrate has to be optimized to get the best transport properties out of the graphene based IC.

Stolyarova et.al.(2007) have taken Scanning Tunneling Microscopy images of graphene and 3-layer graphene on SiO_2. Single layer graphene give honeycomb symmetry, high degree of crystalline order, sub-nanometer fluctuations in height on lateral scale in STM imageries. The integrity of the graphene is maintained inspite of harsh processing condition.FLG show three-fold symmetry.Atomically resolved STM topography show no prominent signs of perturbation induced by interaction with the underlying substrate. Lattice-defects, point-like interaction with the underlying substrate or folding of the single layer graphene sheet would be expected to cause significant perturbation of the local electron density and thus be reflected in STM topography. None of these were observed.

8.3.4.7. Opening the Band-Gap in Graphene. [Ando (2009)]

For mainstream logic applications graphene remains metallic even at Charge Neutral Point or at Dirac Point. This is a major problem in building graphene device based logic gates. But Band-gap can be engineered.

Bandgap of $\Delta E = 0.3$eV can be induced in bilayer graphene but this is more suitable for tunable IR lasers and photo-detectors. Here an electric field is applied perpendicular the device as shown in Figue 25.

The band gap induced = eFd/2 where F = external field due to top gate.

$$n_s = g_s g_v \frac{\Delta^2}{2\pi\hbar^2 v_F^2} \sim 2.5 \times 10^{13} cm^{-3} \tag{23}$$

Δ = interlayer coupling =0.4eV, g_s = spin degeneracy = 2 and g_v = valley degeneracy =2;

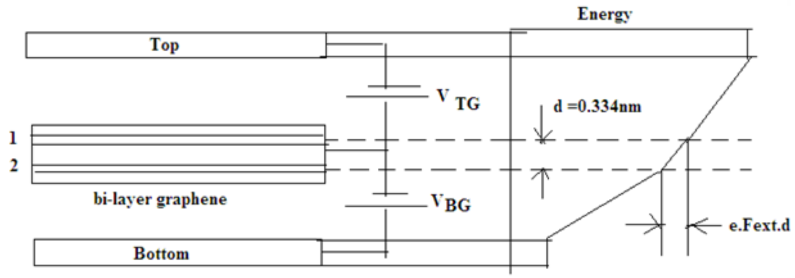


Figure 25. Schematic illustration of bilayer graphene with a bottom gate controlling the carrier concentration and an extra top gate. The distance between layers 1 and 2 is d ≈ 0.334 nm, and the potential diference is determined from eFd. F ext represents the field due to the top gate.

In single layer graphene, band-gap ΔE can be induced by spatial confinement or by lateral-super lattice potential. When graphene is grown over Boron-Nitride or SiC, it has a sizeable gap (>0.1eV) . This is because of large super-lattice effects.

When A and B sub-lattices are at symmetrical energy level then we have energy dispersion given by Eq 1. But if we introduce asymmetry between the sub-lattices say of +δ for sub-lattice A and –δ for sub-lattice B then the energy dispersion relation is given as follows:

$$E_{\pm} = \pm\sqrt{\hbar v_F\,|k|^2 + \delta^2}$$

From Eq.24, at $|k| = 0$, the band gap is 2δ. Such a band-gap is achieved by growing graphene on an appropriate substrate. This lateral super lattice potential of the substrate gives rise to the asymmetry between the two sub lattices.

Functionalizing graphene by reversible hydrogenation can change the electrical property from metallic to semiconducting owing to induced changes of functionalized carbon from sp^2 to sp^3 hybridization[Geim et.al (2007), Gilje et.al.(2007) and Wang et.al(2008)] The resultant hydrocarbon compound, GRAPHANE, can be modified into new materials, fine tuning its electronic property. This has opened up increasingly fertile possibilities in hydrogen storage and in 2D electronics [Berger et.al.(2006),Geim & Novoselov (2007), Geim-Morozov-Hill et.al.(2007)], Gilje et.al.(2007), Novoselov et.al.(2007), Wang et.al.(2008) and Zhang et.al.(2006)]

In Solid State Devices, through potential well confinement band gap can be engineered as is done in quantum dot devices. But in graphene due to Klein Paradox [Section 4.4.3], potential well confinement is not possible. Hence we have to go for spatial confinement.

In Schrodinger Systems, potential well confinement leads to enhanced band gap [Ando (2009)]:

$$\Delta E = \frac{\hbar\pi^2}{2d^2}\left(\frac{1}{m_n}+\frac{1}{m_p}\right) \tag{25}$$

Where d = width of the potential well.

In Schrodinger Systems i.e. in non-relativistic systems, wave function Ψ(x) is zero at x = 0 and at x = d. Therefore electrons are weakly affected by disorder present at the boundary. Hence there is weak boundary roughness scattering.

In graphene we have solutions of Dirac Differential Equation. These solutions are not zero at the edges hence Ψ(x) is quite sensitive to the edge structure.

In zig-zag ribbons, edge states have no dispersion and localized in the vicinity of edges.

In arm-chair ribbons the width decides the conductivity of the ribbon. As width increases the conductivity alternates between that of metal and semi-conductor . This is analogous to CNT.

In actual preparation, the edges are extensively damaged and therefore far from ideal.

Measurement of conductance of narrow ribbon at different temperature show that Band Gap is inversely proportional to the width, it is independent of ribbon direction and conductivity decreases with decreasing ribbon width. Therefore the gap is not band gap but mobility gap related to edge disorder. Chemically derived ribbons may have tunable band-gap.

Experiment no -1: Components Testing
This experiments describes the whole spectrum of active and passive components. The continuity test of these components and their I-V charateristic is described.

Experiment no -1

Name of Experiment :-

Testing of active &passive component with the help of C.R.O and digital multimeter

Aim :- To study the function of digital multimeter & C.R.O & testing various components like passive and active components such as resistance, capacitance , inductance, classical diodes, zener diodes, photodiode, Light Emitting Diodes, BJT, JFET and MOSFET.

Theory:-In component testing process, we test the continuity and the I-V characteristics of various component like resistor, capacitor, diode and transistors.

Wires and Connectors: we use the following wires:

1. Single Strand wire of SWG 22 or SWG23. SWG22 is preferred because it snugly fits the solder-less, spring loaded contact points in the breadboard. SNUGLY means neither loose nor

tight. SWG is Standard Wire Guage. 22 means diameter φ = (1/22) inch. These are used as hookup wire for making connections on the breadboard.

2. Multistrand wires are flexible wires used as connecting wires between the Power supply and the bus of the circuit. Bus are lines which act as GROUND/EARTH, lines which act as (+)ve bus or as (-)ve bus. The guage of multi-strand wire can be SWG 7/36 or SWG 14/36.

SWG 14/36 means 14 strands and 1/36 inch in dia. These multistrand wires have banana plug on one end and crocodile clip on the other. By banana plug they are fitted on the banana socket on the power supply and by crocodile clip it is clipped on the hook up wire connected in the bread board.

Iii There are various types of wires. There are two cores or three cores shielded wires. In one sheath there are two or three insulated SWG14/36 wires. The sheath can be shielded in wire-mesh shields. This prevents Electro-magnetic interference and pick-up. There are unshielded wires with two and three cores.

1. There are coaxial cables which are used as connectors . A coaxial cable with BNC Connector on one end and crocodile clips on the other are used for feeding the signal from Function Generators to the circuit under test. A

coaxial cable with BNC Connector on one end and oscilloscope probes or crocodile clips on the other are used for feeding the signal from the circuit under test to the oscilloscope.

Procedure:-

- Set up the component in the breadboard which has to be tested & probed by the mulimiter as well as C.R.O.
- First of all we calculate the theoretical value of resistor through colour

coding and than obtain practical value through multimeter.

- Next we find the I-V characteristic curve i.e forward and reverse characteristics of diode with the help of C.R.O.

Resistances:-

1. **Carbon Film Reistances.**

There is a carbon film deposited on the former. These are the cheapest and most widely used resistances. The resistances are given in colour codes.

Colour code is Gray, White, Violet ,Blue ,Green, Yellow, Orange, Red Brown and Black. These stand for 9,8,7,6,5,4,3,2,1 and 0 respectively.

The first two bands on the left give tens and units. The third band gives the exponent to the base 10. Fourth band gives the tolerance. Golden Band gives 5% and Silver Band gives 10% tolerance.

S.No	**1st,2nd,3rd and 4th band of colours**	**Nominal Value (Ω)**	**Practical Value (Ω)**	**Tolerance (± %)**	**Percentage Error**
1	Yellow, Violet, Orange, Silver	47k	47.2k	10	0.423
2	Red, Violet, Red, Gold	2.7k	2.748k	5	1.746

The resistances have power rating and tolerance. Power rating tell the maximum power dissipation permitted. These are generally 1/4W or 1/2W rating in carbon film resistances.

The fourth band tells the tolerance. Zero Tolerance resistance is a precision resistance. Less than 1% tolerance is close tolerance component. More than

5% tolerance is a wide tolerance component.

ii. Thin Film Resistances :Thin Film Resistances are made of Nichrome which is 80% Nickel and 20% Chromium alloy. These are of 0.1% tolerance and the temperature coefficient of resistance(t.c.r) is 100 parts per million per degree centigrade (PPM/ºC). These are also called precision resistances.

1. **CERMETS** are film resistances where ceramic and metal films are used. These are also of 0.1% tolerance and zero t.c.r. Ceramic has a negative t.c.r. and Metal has positive t.c.r. Hence a proper combination of Ceramic and Metal gives zero t.c.r. These are truly precision components used in Intrumentation.
2. **Wire Wound Resistances:** These are power resistances which can go up to 100W maximum power dissipation. These are green in colour and robust in appearance.
3. **Variable Resistances:** These are carbon track potentiometers and Rheostats. Carbon track resistances are used in Electronics Lab and are of less than 1W rating. Rheostats are used in Electrical Engineering Labs and can go up to 1000W. Rheostat are made of Cu/Ni or Ni/Cr alloy of low t.c.r.. In Carbon Track Potentiometer there is a carbon track on which a wiper sweeps from one end to the other. The two terminals connected to the two ends of the carbon track give a fixed resistance while the

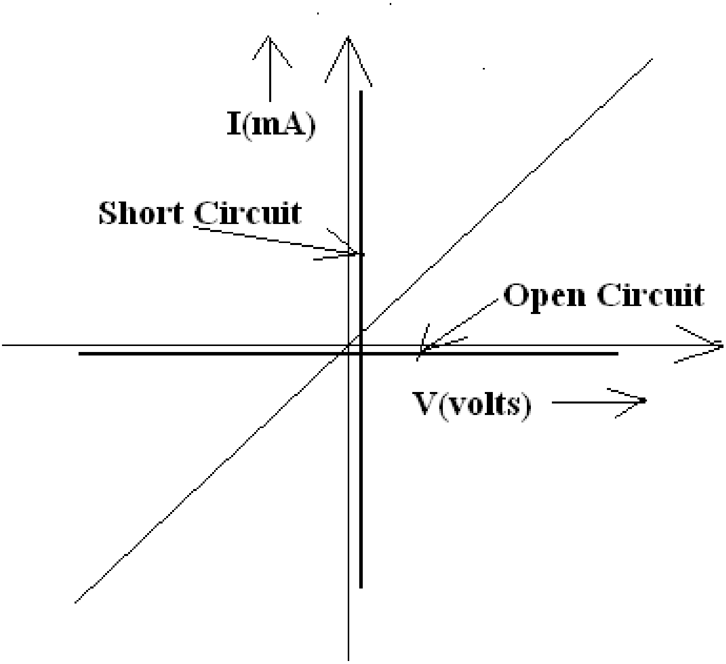
resistance between one of the end terminals and the terminal connected to the wiper is a variable resistance.

Capacitor:

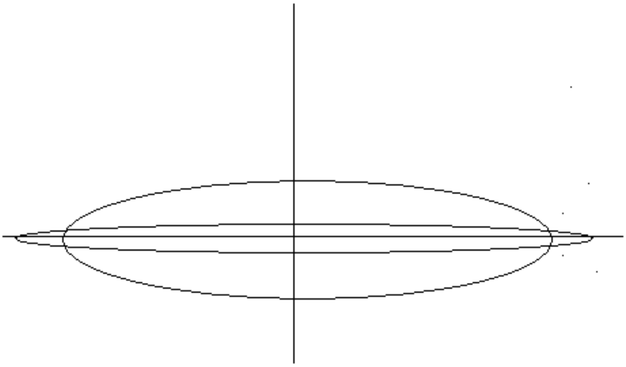
1. **Ceramic Disc Capacitances:**

These are the cheapest and most widely used capacitances. The values are given as 103. 103 means 10×10^3pF (pF is called PUF). This is 10^{-8}F = 0.01μF.

1. **Paper/Mica Capacitances.**
2. **Polyesterene Capacitances:** These are low loss angle high quality capacitances.
3. **Electrolytic Capacitances:** The large magnitude of capacitances more than 1μF up to 1000μF can be realized by electrolytic Capacitances. These are polar capacitances. The polarity of the DC Voltage applied should be correct so that the electrolyte between the electrodes act like dielectric. If the polarity is reversed then it becomes a conductor.
4. **Variable Capacitances:** Gang Capacitances-these are air-gap capacitances used as tuning capacitances in Radio-Broadcast Receivers. Trimmers and Padders are used for alignment of the tuned circuits in Radio Receivers.



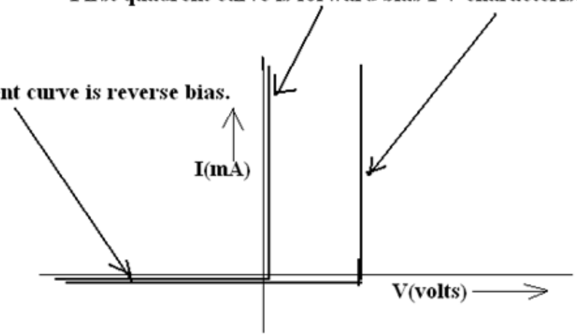
I-V Characteristics of a resistance.



I-V Characteristics of Capacitance.
High eccentricity ellipse is for Ceramic Disk Capacitances
Low eccentricity ellipse is for Electrolytic Capacitance.

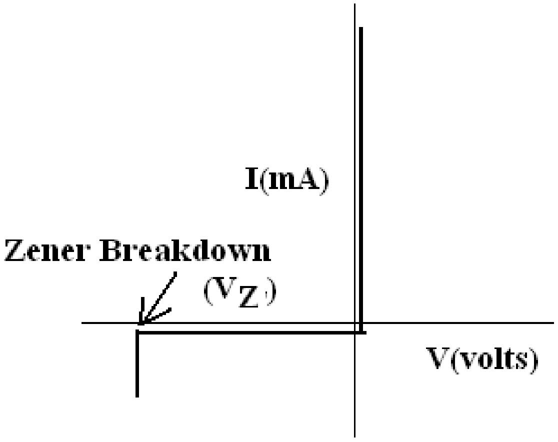
First quadrent curve is forward bias I-V characteristics.

Second quadrent curve is reverse bias.



I-V Characteristics of a classical diode and LED.

In first quadrent we see the cut-in voltage,
Cut-in Voltage is 0v for Ge Diode, it is 0.5V for Si Diode
but it is 1.2V or higher for Compound Semiconductor.
LED are always made of Direct-Band-Gap material and
hence it is made of Compund Semiconductor.



I-V characteristics of Zener Diode.

Zener Diode is Quantum-mechanical Diode.

It is heavily doped.

It experiences Zener Breakdown by Quantum Mechanical Tunneling.

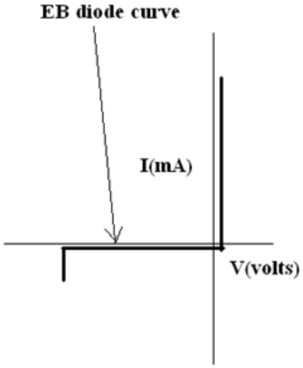
Zener Break Down occurs below 4V.

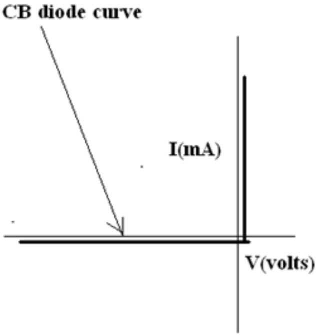
Avalanche Breakdown occurs above 6V to 100V.

Between 4V to 6V there is a combination of Zener Breakdown and Avalanche Breakown.

Hence it can have ZERO tcr.

Zener Breakdown has (-)ve tcr and Avalanche Breakdown has (+)ve tcr.





I-V characteristics between Base and Emitter Diode and between Base and Collector Diode.
Emitter has a doping density of 10^18/cc , Base has 10^17/cc and Collector has 10^16/cc.
EB diode behaves like a Zener Diode because of vey high doping hence its I-V characteristics is like that of a Zener Diode.
BC diode behaves like Classical Diode because of moderate doping hence its I-V characteristics is that of a classical diode.

SSPD-Experiment Number 2
This experiment introduces the students to the Frequency Domain Response and Time Domain Response of RC circuit once configured as Low Pass Filter and once configured as High Pass Filter.

Experiement -2

2.1. Aim

-To study the time domain and frequency domain response of RC circuit which is a first order system .

[The number of independent energy storage elements decide the order of the system. In a parallel or series LC resonance circuit the order is two. The order of a system gives the number of natural frequencies of the system. In first order system there is one natural frequency. This natural frequency decides rise-time/fall-time or the sag transients in TIME DOMAIN RESPONSE while it decides the cut-off frequency/-3dB frequency/0.707 frequency/half power frequency/corner frequency in Frequency Domain Response]

2.2. Apparatus

- digital storage oscilloscope, breadboard, function generator.

2.3 Theory

–

1. The steady state sinusoidal response of the system under consideration is the Frequency Domain Response of the system.

Transfer Function of a system = H(jω) = $V_{out}(j\omega)/V_{in}(j\omega)$.

This has a magnitude as well as a phase. The plot of the magnitude of H(jω) with respect to frequency is the Magnitude Frequency Response and Phase vs Frequency is Phase Frequency Response.

1. The time domain response is the square wave response of the system

2.4. High pass RC circuit (as shown in Figure 1.a.)

The reactance of the series capacitance is given as

$$X_C = \frac{1}{j\omega C}$$

With increase in frequency, the reactance of the series capacitor decreases and therefore the transmission increases with frequency. At very high frequency the capacitive reactance become negligible so the output becomes almost equal to the input and transmission becomes 100%. Since the circuit attenuates the low frequency signals and allows transmission of high frequency signals with little or no attenuation, it is called high pass circuit. This circuit is widely employed as a coupling circuit in RC-coupled Amplifier. This helps isolate the Q-point of the Amplifier by the loading effect of the Source and the Load of the Amplifier.

The voltage transfer function referring to Figure 1.a.

$$H(j\omega) = \frac{V_0(j\omega)}{V_{in}(j\omega)} \qquad 1$$

$$Mesh\,Current = I(j\omega) = \frac{V_{in}(j\omega)}{(R + \frac{1}{j\omega c})}$$

$$V_{out}(j\omega) = I(j\omega) \times R$$

$$= \frac{V_{in}(j\omega)(j\omega c) \times R}{(1 + j\omega c R)}$$

$$\frac{V_{out}(j\omega)}{v_{in}(j\omega)} = H(j\omega) = \frac{\frac{j\omega}{\omega_l}}{1 + \frac{j\omega}{\omega_l}} \qquad 2$$

$$where\ \omega_l = lower\ cut-off\ frequency = lower\ half\ power\ frequency$$
$$= -3dB\ frequency = 0.707\ frequency = Corner\ Frequency = \frac{1}{\tau} = \frac{1}{RC}$$

$$\text{At } \omega \gg \omega_L \qquad H(j\omega) = 1$$

$$At\,\omega=\omega_L,\quad H(j\omega)=\frac{j}{(1+j)}=\frac{1<90degree}{\sqrt{1+1}<45\;degree}=\frac{1}{\sqrt{2}}\angle 45^0=0.707\angle 45^\circ \qquad 3$$

$$At\;\omega \ll \omega_L\,,\; H(j\omega)=\frac{j\omega}{\omega_L}=\frac{\omega}{\omega_L}\measuredangle 90^{\circ} \qquad 4$$

Equation 4 indicates that at low frequency the magnitude of the transfer function increases at +20dB per decade as shown in Figure 1.b.

Figure 1.(i) Magnitude of the transfer function in dB(decibels) vs Frequency(logarithmic scale) of High Pass Filter(HPF)

(ii) Phase of the transfer function in dB(decibels) vs Frequency(logarithmic scale)of HPF.

The two plots together are known as Bode Plot.

Figure 2. The Time domain response of High Pass Filter for three cases: (i) When T $<<\tau$ (ii) When T $>>\tau$ and (iii) when T $\approx\tau$.

From the 10% sag frequency the lower cut-off frequency can be directly calculated.

Suppose the 10% sag frequency is f* then we have the following formulation:

$$f_L=\frac{0.1\times f^{*}}{\pi}=lower\;cut-off\;frequency\;of\;HPF \qquad 5$$

2.5. Low Pass Filter (shown in Figure 3)

Figure 3. RC Circuit configured as Low Pass Filter.

Figure 4. Time Domain Response of Low Pass Filter.

The fourth figure in Figure 4 is for $T << \tau$. In this case we get the mean value only. The square wave is lost.So at very high frequencies, LPF acts as INTEGRATOR.

The frequency domain analysis is given below with reference to Figure 3:

$$\text{The Input Mesh Current} = I(j\omega) = \frac{V_{in}}{R + \frac{1}{j\omega C}} \qquad 6$$

$$V_{out}(j\omega) = I(j\omega) \times \frac{1}{j\omega C} = \frac{V_{in}(j\omega)}{1 + j\omega RC} \qquad 7$$

$$H(j\omega) = \frac{V_{out}(j\omega)}{V_{in}(j\omega)} = \frac{1}{1 + j\omega RC} = \frac{1}{1 + \frac{j\omega}{\omega_H}} \quad \text{where } \omega_H = \frac{1}{RC} = \text{upper cut-off frequency} \qquad 8$$

Figure 5. Frequency Domain Response of LPF.

In Figure 4 for cases $T \sim \tau$, we get RISE TIME and FALL TIME at the LEADING and LAGGING EDGE of the Square Wave.

Rise Time $= t_r =$ time taken to rise from 10% of the peak value of the square wave to 90% of the peak value of the square wave at the leading edge.

Similarly Fall Time = t_f = time taken to fall from 90% of the peak value of the square wave to 10% of the peak value of the square wave at lagging edge.

It can be shown that:

$$t_r \times f_H = 0.35 \qquad 9$$

Thus from Time Domain Studies the frequency Band-Width of a System can be calculated directly without going through the rigour of drawing the Frequency Response Graphs.

APPENDIX: Frequency Spectrum of Infinite Pulse Train.

Figure A. Frequency Spectrum of an Infinite Pulse Train.

In Figure A, in upper half an infinite pulse train is depicted with Pulse Repetition Frequency(PRF) of (1/T) Hz where T= time period of periodicity and τ is the pulse duration and τ/T is duty cycle. In Square wave duty cycle is 50%. In the pulse train shown above T=4τ. As can be seen in the Fourier Series Expansion of the infinite pulse train, the DC spectral component is the mean value ($V_p\tau/T$) and the fundamental frequency component (1/T Hz) and the harmonics spectral component magnitude follows the envelope described by Sinc(θ)= [Sin(θ)/ θ] where θ= [2πn(τ/T)] where n=1,2,3....integers.

The Sinc Envelope experiences the first Zero-cross over at frequency $= 1/\tau$ and second cross-over at $2/\tau$.

This means Fourier Spectral Components has a main lobe of frequency components and side-lobes of frequency components. As long as the System's BW accommodates the main lobe we will get near-faithful reproduction of the pulse train. But if the main lobe gets suppressed then the pulse train will be completely lost.

In LPF, when $T << \tau$ then the fundamental $1/T$ is $>>$ first cross-over frequency hence the main lobe is completely suppressed and the square wave is completely lost. What is left after filtering is the DC component since it is direct-coupled system.

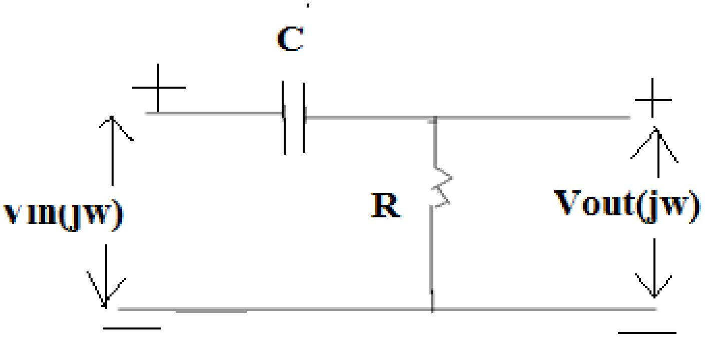
Similarly in HPF, when $T >> \tau$ then the fundamental $1/T << 1/\tau$ and since it is HPF hence the main lobe gets completely suppressed. What is left are the spikes corresponding to the edges of the pulse train. The pulse train itself is completely lost..

SSPD_Experiment2_Frequency Domain Response of HPF

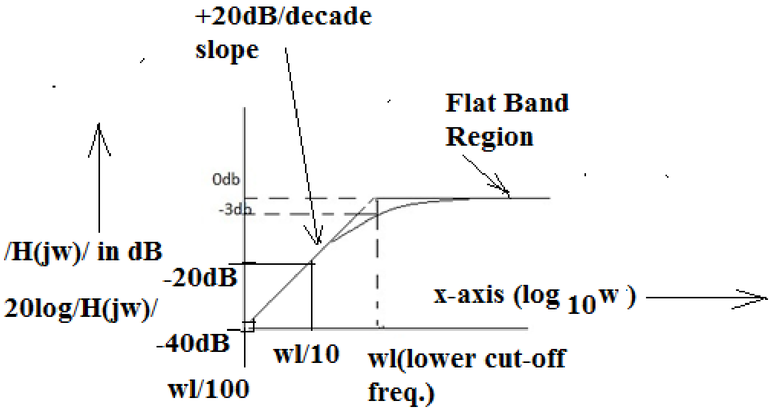
This module gives the BODE PLOT of the Transfer Function of HPF

Experiement -2-Figures of HPF.

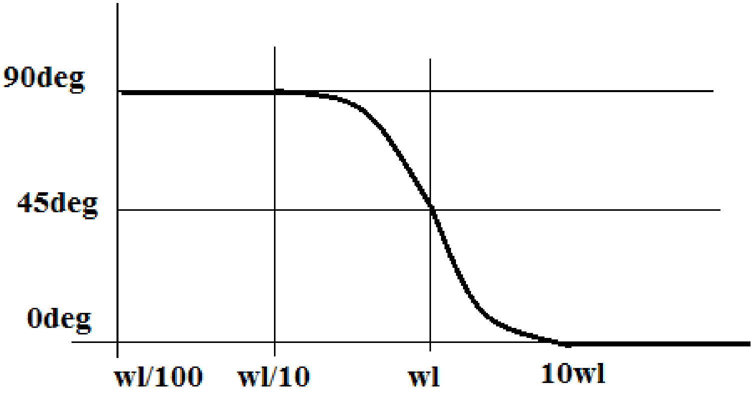
2.4. High pass RC circuit (as shown in Figure 1.a.)



(a) R-C configuration of HPF



(b) Bode Plot of the Transfer Function Magnitude /H (jw)/ with respect to the log to the base 10 of Frequency.



(c) Phase vs Frequency Response of HPF

Figure 1.(a) Circuit Topology of High Pass Filter.

(b) Magnitude of the transfer function in dB(decibels) vs Frequency(logarithmic scale) of High Pass Filter(HPF)

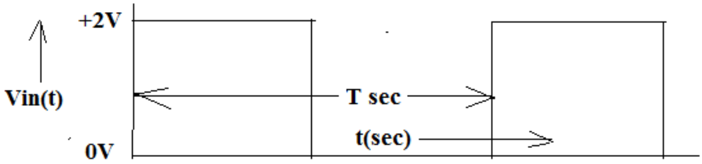
(c) Phase of the transfer function in dB(decibels) vs Frequency(logarithmic scale)of HPF.

The two plots together are known as Bode Plot.

SSPD_Experiment 2_Time Domain Response of HPF. This module gives the Figures of Time Domain Response of HPF.

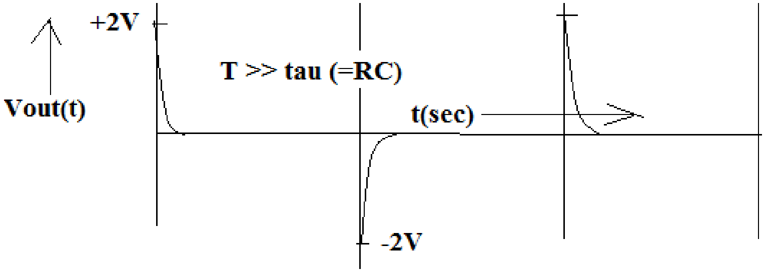
Experiement -2-Time Domain Response of HPF.

2.4. High pass RC circuit .



Input Square Wave of Peak Voltage 2V and an Offset of +1V.The duty cycle is 50%.

Figure 2 .a. Input Square Wave of High Pass Filter .



Square Wave Response of HPF when T >> tau(=RC).

Figure 2 .b. Output Square Wave Response of High Pass Filter for T $>>$ τ .

SSPD_Experiment 2_TDR of HPF_Part2
This module gives the Time Domain Response of HPF for T nearly equal to tau and T much less than tau.

Experiement -2-Time Domain Response of HPF. Part 2

2.4. High pass RC circuit .

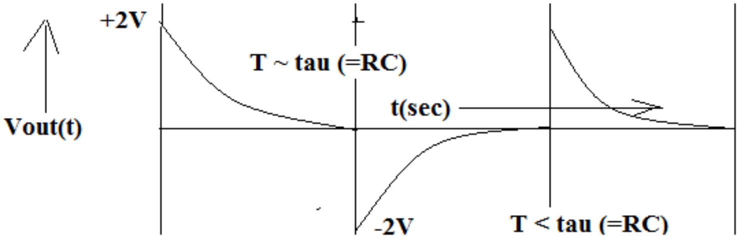


Figure 2 .c. Square Wave Response of High Pass Filter for T ~ τ .

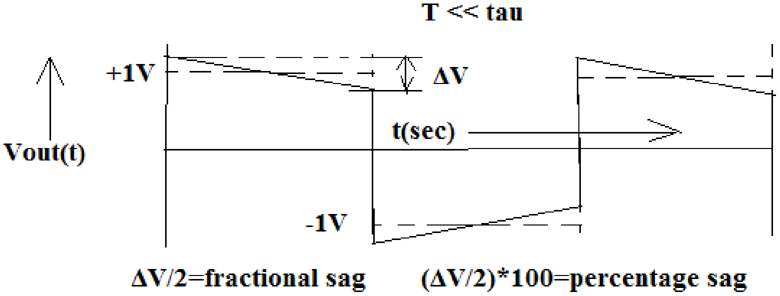


Figure 2 .d. Square Wave Response of High Pass

Filter for $T << \tau$.

When $T << \tau$ then the frequency spectrum of the square wave lies well within the Pass Band of the High Pass Filter. Hence the input square wave is near faithfully transmitted to the output except for a sag at the top and bottom and there is the loss of OFFSET.

The BW of HPF is from $(1/\tau) = \omega_L$ radians per sec to infinity.

The spectral components of the square wave are 0Hz, (1/T)Hz, (3/T)Hz,………

Since $1/T >> (1/(2\pi\tau))$, hence (1/T)Hz, (3/T)Hz, ……… are well accommodated and faithfully transmitted by HPF but 0Hz gets completely suppressed. Hence we get a Square Wave with ZERO mean value and TOP and BOTTOM have sag.

SSPD_Experiment 2_Figures of LPF_Part1
This module gives the Circuit Topology and the Bode Plot of the Magnitude and Phase Angle of the Transfer Function of LPF.

Experiement -2-Figures of LPF. Part 1

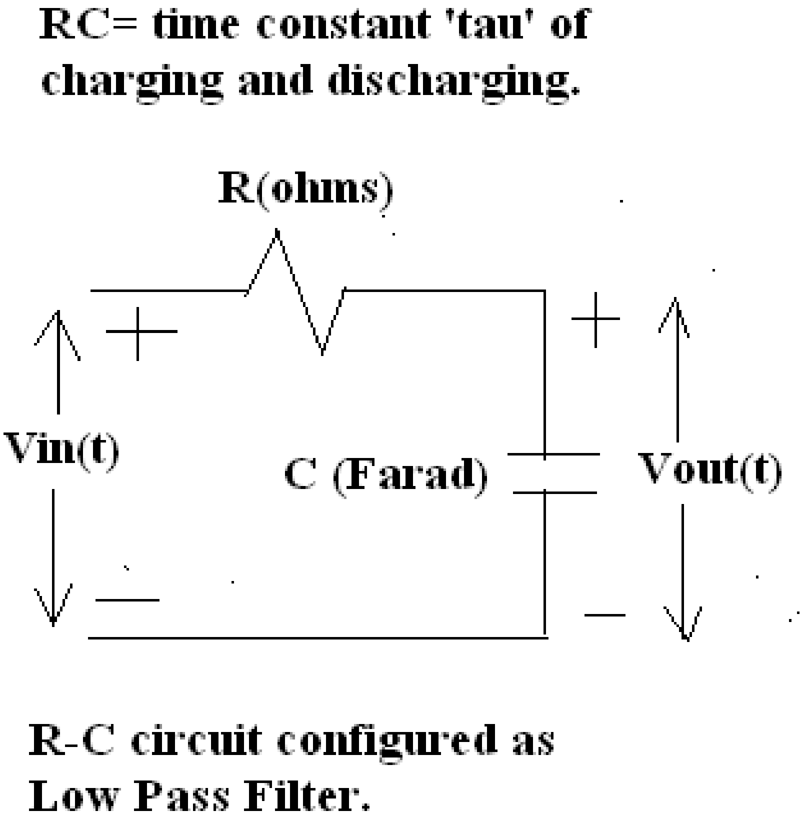
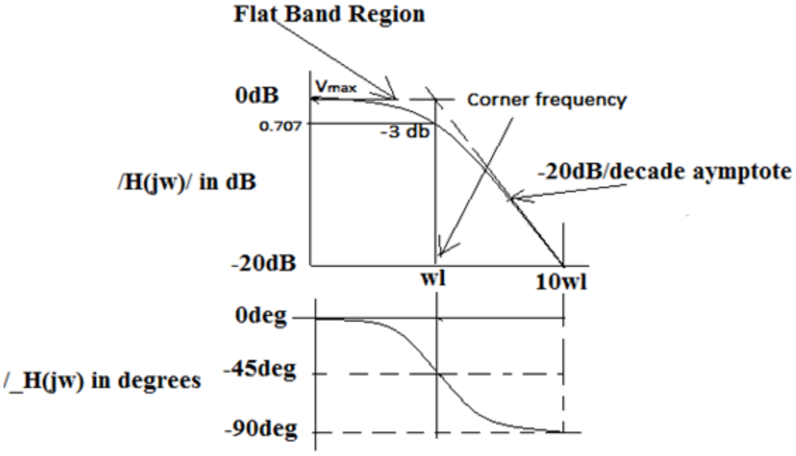

RC= time constant 'tau' of charging and discharging.
R(ohms)
+
Vin(t)
—
C (Farad)
+
Vout(t)
—
R-C circuit configured as Low Pass Filter.

Figure 3. RC Circuit configured as Low Pass Filter.



Bode Plot of Magnitude and Phase Angle of the transfer function H(jw) of LPF.

Figure 5. Frequency Domain Response of LPF.

SSPD_Experiment 2_Time Domain Response of LPF
This module gives the Time Domain Response of LPF.

Experiement - 2 - Figures of TDR of LPF. Part 2.

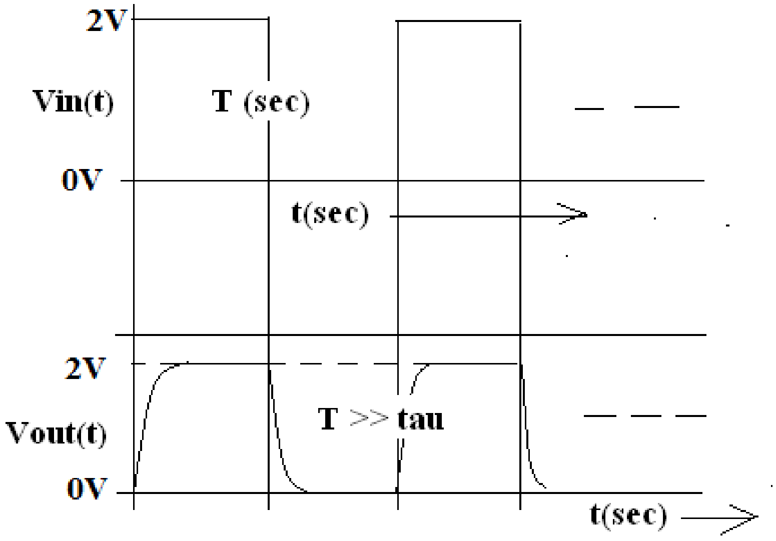


Figure 4 a and b. Input Square Wave and Time Domain Response of Low Pass Filter for T $>>$ τ.

SSPD_Experiment 2_Time Domain Response of LPF_Part2

This module gives the Time Domain Response of LPF when T approximately equal to tau or T is much less than tau.

Experiement - 2 - Figures of TDR of LPF. Part 2.

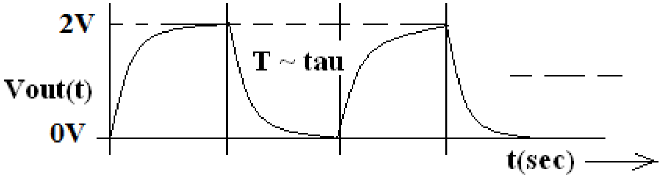
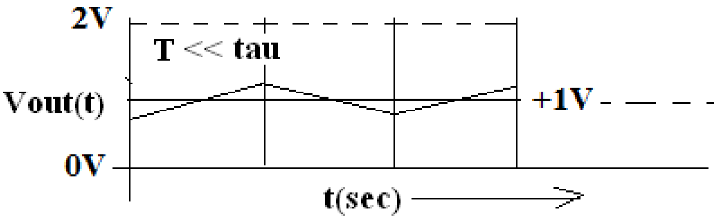


Figure 4 c. Time Domain Response of Low Pass Filter for T ~ τ .

The leading or positive Edge of the Square Wave experiences a Rise Time and the lagging or negative Edge of the Square Wave experiences a fall Time. The suppression of the higher frequency components results into rounding off the EDGES of the Square Wave.



LPF acts as an Integrator for high frequency Square Waves.

Figure 4 d. Time Domain Response of Low Pass Filter for $T << \tau$.

Since (1/T)Hz fundamental component and the harmonics of square wave lie well beyond the Pass Band of LPF hence the nearly all the fundamental and harmonics are suppressed. What is left is the DC component. Therefore at the output we see the Mean Value of the Square wave with a small poitive ramp for (+)ve half and negative ramp for (-)ve half of the Square Wave.

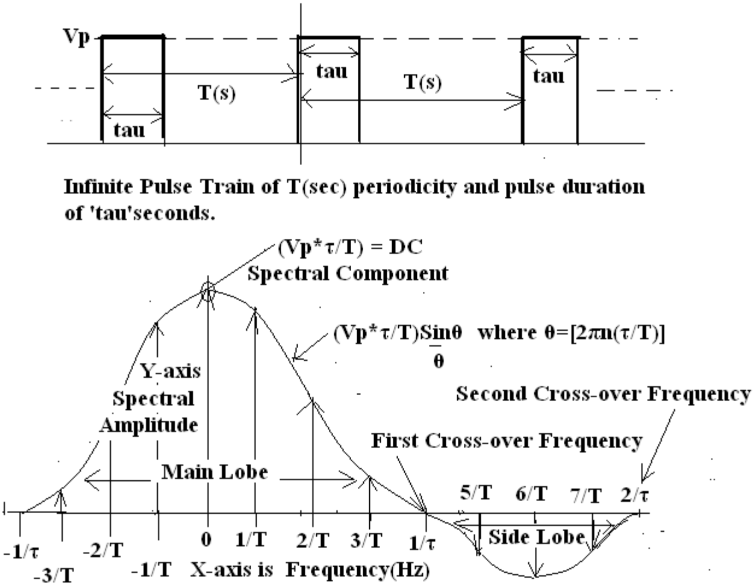

Vp
tau
T(s)
tau
T(s)
tau
Infinite Pulse Train of T(sec) periodicity and pulse duration of 'tau'seconds.
(Vp*τ/T) = DC Spectral Component
(Vp*τ/T)Sinθ/θ where θ=[2πn(τ/T)]
Y-axis Spectral Amplitude
Second Cross-over Frequency
First Cross-over Frequency
Main Lobe
5/T
6/T
7/T
2/τ
Side Lobe
-1/τ
-3/T
-2/T
-1/T
0
1/T
2/T
3/T
1/τ
X-axis is Frequency(Hz)

Figure A. Frequency Spectrum of an Infinite Pulse Train.